



## Datamining für den IceCube-Detektor

Optimierung einer Multiensemblemethode zur Analyse von IC59-Monte-Carlo-Daten

Arbeit zur Erlangung des Bachelorgrades

von

Jan Freiwald

Dortmund, 16. Juli 2012

# **Inhaltsverzeichnis**

<b>1 Einleitung</b>	<b>3</b>
<b>2 Kosmische Strahlung</b>	<b>4</b>
2.1 Atmosphärische Neutrinos . . . . .	5
<b>3 Der IceCube-Detektor</b>	<b>6</b>
<b>4 Verfahren der Datenanalyse</b>	<b>6</b>
4.1 Kreuzvalidierung . . . . .	8
4.2 Random Forest . . . . .	9
<b>5 Analyse der IC59-Daten</b>	<b>10</b>
5.1 Clustering . . . . .	11
5.2 Beschreibung des Separationsprozesses . . . . .	12
5.3 Ergebnis des Gesamtprozesses . . . . .	15
<b>6 Resümee und Ausblick</b>	<b>16</b>
<b>7 Anhang</b>	<b>18</b>
<b>Literatur</b>	<b>57</b>

# 1 Einleitung

Ein wichtiger Bestandteil der Physik ist die Beobachtung der Natur. Diese Beobachtungen beschränken sich stets auf das Mess- und Nachweisbare. [1]

Die Astrophysik stellt die Fragen nach den Mechanismen, dem Ursprung und der Natur des Kosmos. Sie stützt sich dabei auf eben diese Beobachtungen, mit deren Hilfe Theorien unterstützt oder widerlegt werden können. Durch die Erforschung der Teilchenphysik wird das Spektrum der nachweisbaren Partikel größer und es lassen sich mehr Teilchen als nur Photonen, die seit jeher Quelle für astronomische Messungen sind, detektieren und nutzen.

Neutrinos, die sich durch ihren sehr geringen Wechselwirkungsquerschnitt in der Größenordnung von  $10^{-37} \text{ cm}^2 \cdot E_\nu/\text{GeV}$  (für 10 GeV Neutrinos) [2] und ihre neutrale elektrische Ladung auszeichnen, bewegen sich geradlinig fort und werden nur selten absorbiert oder gestreut. Sie sind ein Kandidat für die Erforschung weit entfernter astronomischer Objekte. Neutrinos können mit Materie wechselwirken und dabei in ihre leptonischen Partner umgewandelt werden, welche sich wiederum über Cherenkovlicht detektieren lassen. Der IceCube-Detektor kann dieses Licht messen. Daraus lassen sich Energie und Winkel des leptonischen Partners rekonstruieren. Mit diesen Daten können dann Rückschlüsse auf Zerfälle angestellt werden.

Ziel dieser Arbeit ist die Trennung von atmosphärischen Neutrinos und Hintergrundereignissen. Dazu soll eine Multiensemblemethode benutzt werden, die den Random Forest-Algorithmus<sup>1</sup> und eine Vorsortierung, das Clustering, auf simulierte Ereignisse anwendet. Dabei soll die Vorsortierung so vorgenommen werden, dass physikalisch ähnliche Ereignisse zusammengefasst werden. Durch eine Gruppierung der Daten lässt sich der Algorithmus auf gruppeninhärenten Strukturen sensibilisieren. Durch eine Rekombination der Gruppen kann die Erkennungsrate und Genauigkeit des Algorithmus gesteigert werden. Das ist nötig, um eine größere Statistik – besonders in höheren Energiebereichen – für die Rekonstruktion eines Neutrinospektrums zu erhalten und damit den Fehler dieser Rekonstruktion geringer zu machen. [4]

In Abschnitt 2 werden die kosmische Strahlung und der Ursprung der Neutrinos erklärt. Abschnitt 3 beschreibt kurz den Aufbau des IceCube-Detektors. Daraufhin werden in Abschnitt 4 die zur Analyse verwendeten Verfahren vorgestellt und erläutert. Der Analyseprozess wird in Abschnitt 5 beschrieben. Zum Schluss wird in Abschnitt 6 das Ergebnis diskutiert.

---

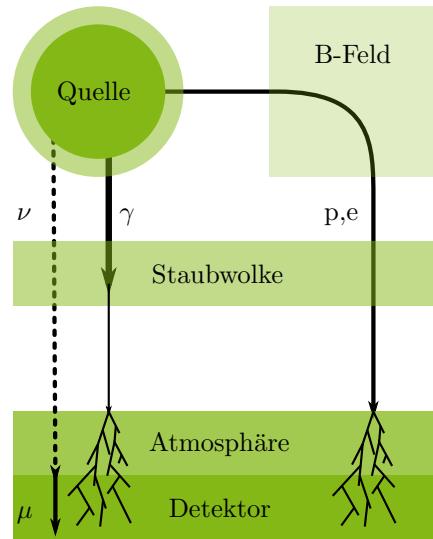
<sup>1</sup>Der Random Forest eignet sich besonders zur Untersuchung von großen Datenmengen mit vielen Attributen [3].

## 2 Kosmische Strahlung

Teilchen unterliegen, nachdem sie von einer Quelle, beispielsweise einem aktiven galaktischen Kern, ausgesandt wurden, je nach ihren Eigenschaften, unterschiedlichen Prozessen auf ihrem Weg zur Erde. Sind die Teilchen geladen, können sie durch magnetische Felder abgelenkt werden und dadurch auf Umwegen zur Erde gelangen. Sie verlieren dabei jegliche Information über den Ort ihrer Herkunft, was sie zur Lokalisation und Untersuchung von Quellen ungeeignet macht. [5]

Photonen stellen eine bessere Basis für astronomische Untersuchungen dar, weil sie nicht durch besagte Felder abgelenkt werden und mit einfachen Mitteln messbar sind. Allerdings werden sie durch Staub abgeschirmt, der die Quelle umgeben oder sich im interstellaren Raum befinden kann. [6] Zudem wird das Universum für hochenergetische Photonen ab 200 TeV undurchsichtig, da Wechselwirkungen  $\gamma\gamma_{3K} \rightarrow e^+e^-$  mit dem 3K-Mikrowellenhintergrund stattfinden [7]. IceCube hingegen erlaubt es Neutrinos mit Energien bis in den PeV-Bereich zu messen. [8] Besonders geeignet zur Erforschung von Punktquellen scheinen Neutrinos zu sein, da sie sich auf einer geraden Bahn bewegen und nur schwach und mit extrem geringem Wirkungsquerschnitt wechselwirken. Sie zeigen sich also unbeeinflusst von Staub oder Magnetfeldern. Allerdings sind es gerade diese Eigenschaften, die zu Schwierigkeiten bei der Detektion von Neutrinos führen. [6] Als Quellen von kosmischen Neutrinos kommen verschiedene Objekte in Frage [9] [10]:

**Ein aktiver galaktischer Kern** ist ein supermassives schwarzes Loch im Inneren einer Galaxie. Um dieses Objekt entsteht eine Akkretionsscheibe, die im optischen Bereich besonders stark strahlt. Zudem bilden sich hoch relativistische Jets senkrecht zu eben jener Scheibe. [10]



**Abbildung 1:** Die Darstellung zeigt stark vereinfacht die unterschiedlichen Prozesse, die auf die Teilchen wirken. Neutrinos zeigen sich von Störfeldern unbeeinflusst, wohingegen geladene Teilchen abgelenkt und Photonen in vergleichsweise hohem Maße absorbiert werden. (Abb. nach [6])

**Gamma-Ray-Bursts** sind starke Gamma-Strahlungs-Quellen, die zirka einmal pro Tag auftreten. Sie können durch ihre räumlich isotrope Verteilung als extragalaktisch angenommen werden. GRB's strahlen so stark, dass sie durch ihre hohe Leuchtkraft andere Himmelskörper überstrahlen können. Gamma-Ray-Bursts sind einige Sekunden lang und produzieren Photonen, die große Energiebereiche abdecken. [9]

Das Spektrum für kosmische Neutrinos wird proportional zu  $E^{-2}$  vorhergesagt. [11]

## 2.1 Atmosphärische Neutrinos

Atmosphärische Neutrinos entstehen hauptsächlich wenn geladene Pionen in der Erdatmosphäre wechselwirken. Diese entstehen hauptsächlich durch Nukleon-Nukleon, Nukleon-Gamma Reaktionen und den Zerfall von Kaonen. [12] Der Zerfall eines positiv geladenen Pions

$$\pi^+ \longrightarrow \mu^+ + \nu_\mu \quad (1)$$

produziert ein Myon-Neutrino und ein Myon, welches anschließend mit der Reaktion

$$\mu^+ \longrightarrow e^+ + \nu_e + \bar{\nu}_\mu \quad (2)$$

zerfallen kann oder im Detektor als Untergrund gemessen wird. Dieser lässt sich allerdings durch die Erde abschirmen, da Myonen nicht weit genug durch die Erde propagieren können. Der gleiche Prozess findet ladungskonjugiert auch für das  $\pi^-$  statt. Innerhalb von Materie können die so entstandenen atmosphärischen Neutrinos mit Nukleonen wechselwirken, wobei sie, unter Leptonen-Familienzahl-Erhaltung, in das entsprechende Lepton und einen hadronischen Rest übergehen:

$$\nu_x + N \longrightarrow l_x + X \quad (3)$$

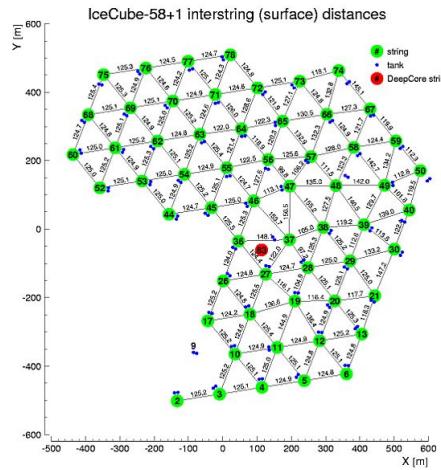
Das  $\nu_x$  steht dabei für  $e$ -,  $\mu$ - und  $\tau$ -Neutrinos. Diese können durch Oszillationen ineinander übergehen. Auf der Erde wird ein Verhältnis von  $1 : 1 : 1$  vorhergesagt, das aus dem Verhältnis  $2 : 1 : 0$  an der Quelle durch Neutrino-Oszillation zustande kommt. [12]

Neutrinos können auch durch den Zerfall von charmhaltigen Mesonen entstehen. Diese Zerfälle werden allerdings erst ab Energien von zirka  $5 \cdot 10^7$  GeV relevant. Durch diese Prozesse entstandene Neutrinos werden als prompte Neutrinos bezeichnet. [12] Das Spektrum der atmosphärischen Neutrinos ist proportional zu  $E^{-3.7}$ . [13]

### 3 Der IceCube-Detektor

Der IceCube-Detektor befindet sich am geographischen Südpol. Hauptbestandteil des Detektors ist ein dreidimensionales Gitter aus digitalen optischen Modulen (DOMs), die in einer Tiefe von 1450 m bis 2450 m im antarktischen Eis eingeschmolzen sind. Jedes dieser Module hat einen Durchmesser von zirka 25 cm. Sie sind mit Photomultipliern (PMT) zur Detektion von Cherenkovlicht ausgestattet. Die Elektronik wird durch eine Glaskugel geschützt. Die DOMs sind in regelmäßigen Abständen von 17 m zu 60 Stück an einem sogenannten String angebracht. Insgesamt 86 Strings sind in einem triangulären Gitter mit einem Abstand von etwa 125 m im Eis eingelassen, sodass ein Gesamtdetektorvolumen von  $1 \text{ km}^3$  erreicht wird. Die Ausleseelektronik kommuniziert über die Strings mit der Oberfläche. [8]

Der zur Analyse verwendete Datensatz bezieht sich auf eine Detektorgeometrie in einem früheren Baustadium. Im Jahr 2009 waren erst 59 Strings vorhanden, so dass ein Kreisbogenausschnitt des beschriebenen Detektors nicht vorhanden ist. [14]



**Abbildung 2:** Die Darstellung zeigt eine Aufsicht der IceCube-59-Geometrie aus dem Jahr 2009. [14]

### 4 Verfahren der Datenanalyse

Zur Separation der IC59 Monte-Carlo-Daten, in Signal (*true*) und Untergrund (*false*), soll ein Klassifizierungsverfahren des maschinellen Lernens angewandt werden. Dazu sol-

len zur Bildung eines geeigneten Modells Monte-Carlo Simulationen verwendet werden, da bei simulierten Daten bekannt ist, ob ein Ereignis Signal oder Untergrund darstellt. Als Signal gelten die atmosphärischen Neutrinos aus Abschnitt 2. Die Nebenprodukte der Zerfälle aus diesem Abschnitt, die Cherenkovstrahlung auslösen können, stellen den Untergrund dar.

Zwei für die Analyse wichtige Größen sind die Precision (bei zwei Klassen auch Purity) und der Recall (bei zwei Klassen auch Efficiency) eines Modells. Diese werden aus der Anzahl der richtig klassifizierten Signale  $t_p$ , der Anzahl der falsch als Untergrund klassifizierten Signale  $f_n$  und den falsch als Signal erkannten Untergrundereignissen  $f_p$  berechnet. Die Precision

$$P = \frac{t_p}{t_p + f_p} \quad (4)$$

beschreibt die Genauigkeit des Klassifikators. Der Recall ist durch

$$R = \frac{t_p}{t_p + f_n} = \frac{t_p}{t_{\text{ges}}} \quad (5)$$

gegeben und macht eine Aussage über die Trefferquote des Modells. [3]

## 4.1 Kreuzvalidierung

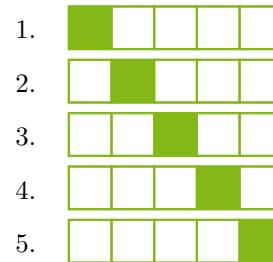
Die Kreuzvalidierung ist ein Verfahren zur Abschätzung der Güte eines Modells. Dabei wird zunächst ein Exampleset in  $n$  gleich große, disjunkte Teilmengen  $T_i$  aufgespalten. Im zweiten Schritt der Kreuzvalidierung wird für jedes  $T_i$  die Restmenge ermittelt. Mit dieser Restmenge wird ein Lernalgorithmus trainiert und ein Modell gebildet.

Anschließend wird das Modell auf  $T_i$  angewendet. Dieser Prozess ist exemplarisch in Abbildung 3 illustriert.

Da bekannt ist, welche Klasse ein Example aus  $T_i$  hat, lässt sich für jedes prüfen, ob es bei der Anwendung des Modells richtig klassifiziert wurde. Dadurch lassen sich Größen, wie der Fehler der Precision oder des Recalls des Modells, bestimmen. [15]

Im letzten Schritt der Kreuzvalidierung

werden der Mittelwert und der Fehler des Mittelwerts für jede ermittelte Größe berechnet. Durch die Größe des Fehlers lässt sich abschätzen, wie stabil der Algorithmus ist, der zur Modellbildung verwendet wurde. Dieser statistische Fehler ist also wichtig zur Einschätzung der Ergebnisse. [15]



**Abbildung 3:** Schematische Darstellung (nach [15]) der Kreuzvalidierung für  $n = 5$ . In jedem Schritt wird auf der gefüllten Menge getestet und auf den Übrigen gelernt.

## 4.2 Random Forest

Ein Random Forest Algorithmus kann zur Klassifizierung von Ereignissen genutzt werden. Wie der Name des Algorithmus impliziert, ist der Random Forest ein Ensemble aus  $I$  verschiedenen Entscheidungsbäumen (siehe [16], vgl. Abbildung 4). An jedem Knoten werden zufällig  $K$

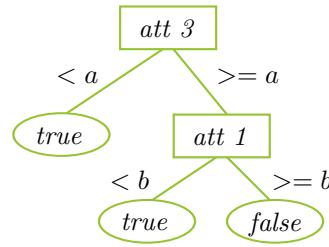
Attribute aus der Gesamtheit gezogen.

Aus diesen wählt jeder dieser Entscheidungsbäume ein Attribut aus und macht einen Schnitt auf ihnen. Wie dieses Attribut und der Schnitt genau gewählt werden hängt von der Implementierung des Entscheidungsbaum-Algorithmus ab und soll hier nicht weiter erläutert werden. [17]

Beim Anwenden des so generierten Modells geben alle Entscheidungsbäume für ein nicht klassifiziertes Example ein Votum über dessen Klasse ab. An dieser Stelle wird davon ausgegangen, dass nur die Klassen *true* und *false* existieren. Dadurch kann die Signalness

$$S = \frac{\text{num}(\text{true})}{I} \in [0, 1] \quad (6)$$

berechnet werden, wobei  $\text{num}(\text{true})$  die Anzahl der Entscheidungen für *true* ist. Ein Ereignis wird genau dann als Signal erkannt, wenn gilt  $S >= S_t$ , wobei  $S_t$  als Signalnesscut bezeichnet wird. Bei einem Signalnesscut von 1 müssen also alle Bäume das Ereignis als Signal klassifiziert haben, damit das Modell eine Aussage über die Klasse (Signal oder Untergrund) machen kann. Welche Werte von  $I$  und  $K$  zu einer dem Problem möglichst angepassten Klassifizierung führen, lässt sich a priori nicht beantworten, sondern ist von der Datenmenge, aus welcher der Random Forest gebildet wird und von jener, auf der das Modell letztlich angewandt wird, abhängig. [3]



**Abbildung 4:** Schematische Darstellung eines Entscheidungbaumes, der aus den Attributen *att 1-3* generiert wurde. Für *att 1* wird bei *b* geschnitten, für *att 3* ist *a* der Schnitt. Das Attribut *att 2* wurde vom Algorithmus nicht gewählt.

## 5 Analyse der IC59-Daten

Es wurden Daten verwendet, die mit NuGen [18] und CORSIKA [19] produziert wurden. Dabei simuliert das Monte-Carlo-Programm NuGen Neutrinos (true) und CORSIKA, das ebenfalls ein Monte-Carlo-Programm ist, den Hintergrund (false). Es wurden im Vorfeld zwei harte Schnitte<sup>2</sup> <sup>3</sup>

$$\text{LineFit\_FVel} > 0.19 \text{ und} \quad (7)$$

$$\text{MPEFit\_Zenith} > 1.54 \text{ rad} \quad (8)$$

auf die Daten angewandt. Hart bedeutet, dass die Ereignisse, welche die Bedingungen der Schnitte nicht erfüllen gelöscht werden. Beide Schnitte (aus [3]) verwerfen wesentlich mehr Hintergrund als Signal und gerade durch den Zenit-Cut werden viele Untergrundereignisse bereinigt, denn die Erde fungiert als Abschirmung (Für eine Erklärung der Schnitte siehe [3]).

---

<sup>2</sup>LineFit\_FVel beschreibt die Geschwindigkeit eines Teilchens im Detektor. [3]

<sup>3</sup>MPEFit\_Zenith beschreibt den mit der MPE-Methode errechneten Zenitwinkel eines Ereignisses im Detektor. [4]

## 5.1 Clustering

Durch Gruppierung der Trainingsdaten kann der Random Forest auf die spezifischen Muster einer Ereignisgruppe sensibilisiert werden und Signaturen, die zu dieser Gruppe gehören, besser unterscheiden. Ein Teilchen, das von oben durch den Detektor geht und nur einen String anspricht, ist einer anderen Ereignisgruppe zugehörig, als ein Teilchen, welches quer durch den Detektor propagiert und von vielen Strings gemessen wird. Da diese beiden Arten sowohl in Signal- als auch in Untergrunddaten vorkommen, kann durch eine Gruppierung die Relevanz anderer Attribute erhöht und die Separation damit verbessert werden. [4]

Die Daten werden an den in Tabelle 1 angegebenen Punkten geschnitten. Für jede dadurch entstandene Gruppe wird ein Random Forest trainiert. Eine Darstellung der geometrischen Cluster findet sich in Abbildung 5.

Die Entfernung des Zerfallsvertex des Neutrinos von der Mittelachse des Detektors ist durch

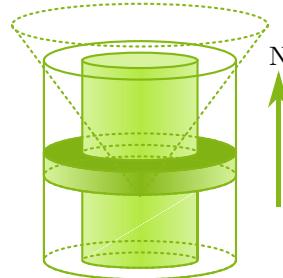
$$\text{vertex\_radius} = \sqrt{r_x^2 + r_y^2}$$

gegeben. Je näher er an dieser liegt, desto höher ist die Wahrscheinlichkeit, dass die gesamte Energie im Detektor deponiert wird und die Rekonstruktionsalgorithmen

durch mehr Informationen über das Ereignis genauere Werte liefern. [4] Das Attribut wurde in zwei Teilmengen zerlegt: Eine mit Vertexpositionen nahe der Hauptachse und eine dazu disjunkte Menge mit weiter entfernten Vertices.

Der Zenitwinkel gibt an in welchem Winkel ein rekonstruiertes Teilchen auf den Detektor trifft. Dabei entsprechen 0 rad einem Teilchen direkt aus dem Himmel über IceCube und ein Ereignis mit 3.14 rad hätte seinen Ursprung beispielsweise im Erdmittelpunkt. [4] Der Schnitt bei 2.35 rad wurde nach einer Monte-Carlo-Studie gewählt, es besteht die Möglichkeit, dass er zu einer besseren Trennung führt. Die Verteilung des Zenitwinkels ist in Abbildung 10 dargestellt.

Zur Rekonstruktion eines Ereignisses lassen sich verschiedene Methoden verwenden. Die Einfachste ist die Methode der kleinsten Fehlerquadrate. Allerdings stellt die Log-



**Abbildung 5:** Die Skizze zeigt den quazi-zylindrischen IceCube-Detektor mit eingezeichneten, nicht maßstabsgetreuen geometrischen Clustern. Entlang der z-Achse wird auf COGZ und in radialer Richtung auf dem Abstand des Vertex zur z-Achse geclustert. Der Zenitwinkel wird ebenfalls in zwei Bereiche unterteilt.

Attribut	Beschreibung	Schnitt(e) bei
vertex_radius	Entfernung des Vertex	400 m
MPEFit_Zenith	Zenit Winkel	2.35 rad
MPEFit_rlogl	Wert der reduzierten Log-Likelihood-Funktion	9.5
COGZ	Z-Schwerpunkt des Ereignisses	-150 m 50 m

**Tabelle 1:** Die Tabelle zeigt die Werte der Schnitte für das Clustering. Das Attribut COGZ wird in 3 Teile geteilt, die anderen in jeweils 2 Teilmengen.

Likelihood-Methode (vgl. [20]) eine bessere Alternative dar. Dabei wird der Wert der Log-Likelihood-Funktion minimiert. Wird dieser anschließend noch auf die Anzahl der Freiheitsgrade normiert. Daraus ergibt sich der Wert der reduzierten Log-Likelihood-Funktion

$$RLogL = \frac{\log \mathcal{L}}{N_{ch} - 5}. [21]$$

Dieser Wert ist ein Qualitätsparameter der eine Aussage über die Güte des Fits macht. Ereignisrekonstruktionen mit kleinen Werten haben eine hohe Güte und damit mehr Aussagekraft als die mit großen Werten. Hier wird nach einer Untersuchung der Signal- und Untergrundverteilung bei  $RLogL = 9.5$  geschnitten. Erwartungsgemäß sollte der Cluster mit kleineren Werten bessere Ergebnisse bezüglich Recall und Precision liefern. Die Abbildung 9 zeigt Histogramme für  $RLogL$ .

Das Attribut COGZ entspricht der z-Koordinate des Schwerpunktes eines Ereignisses. [4] Das Eis, welches beim IceCube-Detektor als Medium für die Ausbreitung von Cherenkov-Licht genutzt wird, weist zwischen zirka -150 m und 50 m eine Verunreinigung durch Staub auf. Diese Schicht wird als Dustlayer bezeichnet. Der Nullpunkt der Z-Achse befindet sich in der Mitte des IceCube-Arrays. Das Attribut COGZ wurde so geschnitten, dass 3 Gruppen entstehen: Eine Gruppe hat ihren Schwerpunkt oberhalb, die Zweite innerhalb und die Letzte unterhalb des Dustlayers. In Abbildung 7 lässt sich an der Position des Dustlayers ein deutlicher Einbruch im Histogramm des Z-Schwerpunktes der NuGen Ereignisse erkennen. [22]

## 5.2 Beschreibung des Separationsprozesses

Der Separationsprozess ist prinzipiell in drei Schritte geteilt:

**Eine Vorgruppierung** der Daten, wie in Abschnitt 5.1 erklärt.

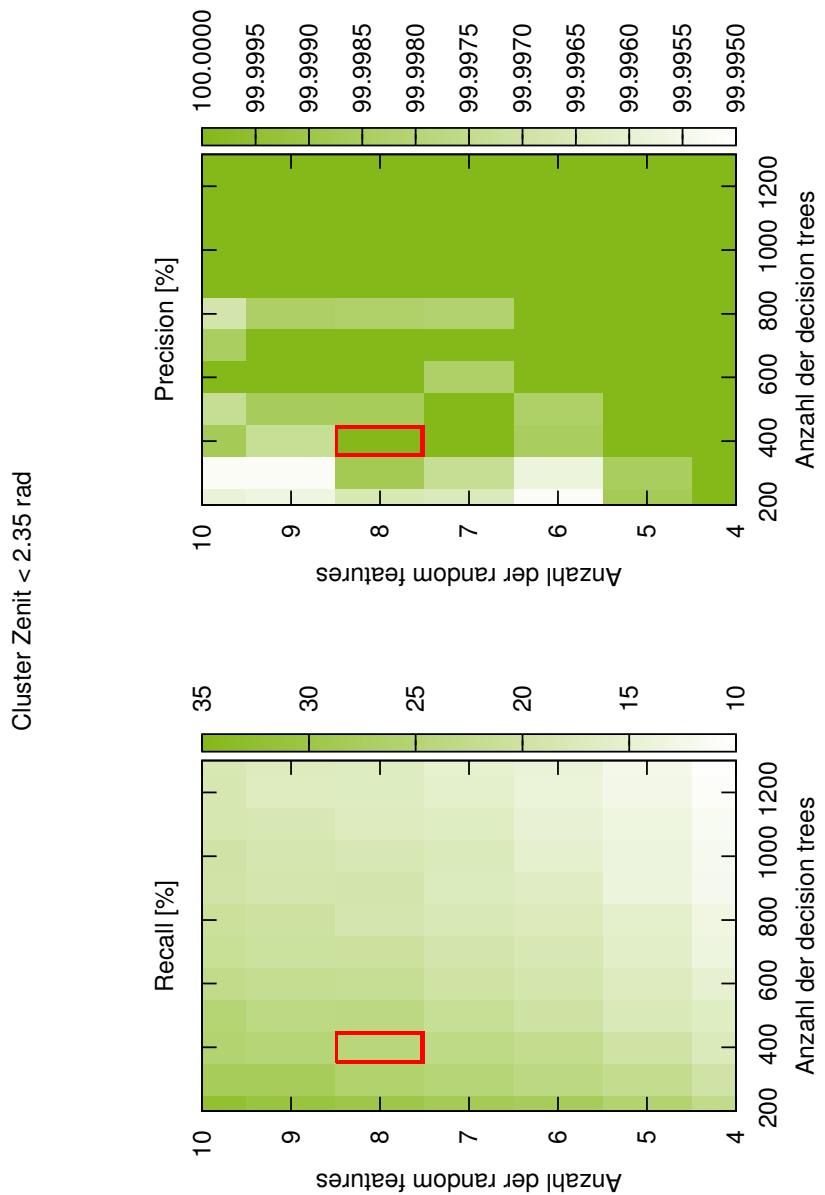
**Die Clusteroptimierung** entspricht der Auswahl geeigneter Random Forest Parameter im Parameterraum von  $I$  und  $K$  für jeden einzelnen Cluster.

**Der Gesamtprozess** beinhaltet die Separation der Daten mit den ermittelten Parametern.

Für die in Abschnitt 5.1 beschriebenen Cluster ist a priori nicht bekannt, welche Parametereinstellungen des Random Forest gewählt werden sollten. Daher werden für die Parameter  $I$  und  $K$  unterschiedliche Einstellungen getestet. Um den Prozess zu vereinfachen und auf Grund von vorhergegangenen Analysen ([3] und [4]) wurde die Anzahl der Trainingsexamples auf je 25000 Signal- und Untergrundereignisse festgelegt. Für die Optimierung der Cluster werden insgesamt 600000 Examples mit gleichen Teilen an Signal und Untergrund verwendet. Bei der verwendeten Kreuzvalidierung handelt es sich um eine 5-fache Kreuzvalidierung. Diese wurde im Rahmen des trade off von Rechenzeit und Genauigkeit gewählt. Es wurde ein Signalnesscut von 1 verwendet. Für die Einstellungen von  $I = 100$  bis 1300 in Schritten von 100 und  $K = 4$  bis 9 in Schritten von 1, werden die Größen  $t_p$  und  $f_p$ , sowie deren Fehler durch die Kreuzvalidierung ermittelt. Innerhalb dieses Parameterraumes lassen sich gute Werte bezüglich Precision und Recall und in Anbetracht des oben erwähnten trade off erreichen. [4] [3]

Es zeigt sich deutlich, dass  $t_p$  und  $f_p$ , und damit auch Recall und Precision, von  $I$  und  $K$  abhängen (siehe Abbildung 6). Für wenige Entscheidungsbäume wird ein größerer Recall erreicht, allerdings nimmt dabei die Precision ab. Analog wird für viele random Features der Recall größer und die Precision geringer. Die Einstellung, die letztlich verwendet wird, wurde so gewählt, dass zunächst  $f_p$  klein und  $t_p$  an zweiter Stelle möglichst groß wird, um eine möglichst hohe Precision mit dem Gesamtprozess zu erreichen.

Im Gesamtprozess werden die einzeln optimierten Random Forests auf 50000 Examples trainiert. Getestet wird auf 380000 Ereignissen. Wenn einer der Cluster ein Ereignis als Signal erkennt, wird es im Weiteren als solches behandelt, unabhängig davon, ob ein anderer Cluster es als solches wertet. Anschließend wird aus diesen Daten wieder Recall und Precision berechnet. Der Fehler, der dabei gemacht wird, lässt sich über die Summe der Fehlerquadrate der einzelnen kreuzvalidierten Cluster nach oben abschätzen. Für den Gesamtprozess wurden die in Tabelle 2 aufgelisteten Parameter verwendet.



**Abbildung 6:** Der Plot zeigt Precision und Recall für den Cluster Zenitwinkel < 2.35. Recall und Precision hängen deutlich von  $I$  und  $K$  ab. Der Plot ist farblich codiert. Je dunkler, desto größer sind Precision oder Recall. Das rote Rechteck zeigt, welche Einstellungen für den Gesamtprozess verwendet wurden.

Cluster	<i>I</i>	<i>K</i>	<i>R</i> [%]	<i>P</i> [%]
vertex_radius < 400 m	400	9	$28.83 \pm 0.59$	$99.999 \pm 0.001$
vertex_radius > 400 m	400	9	$25.09 \pm 0.09$	$100.000 \pm 0.000$
MPEFit_Zenith < 2.35 rad	400	8	$24.37 \pm 0.68$	$100.000 \pm 0.000$
MPEFit_Zenith > 2.35 rad	400	8	$11.51 \pm 0.40$	$100.000 \pm 0.000$
MPEFit_rlogl < 9.5	300	5	$33.52 \pm 0.47$	$100.000 \pm 0.000$
MPEFit_rlogl > 9.5	400	5	$3.64 \pm 0.14$	$100.000 \pm 0.000$
COGZ < -150 m	200	6	$20.03 \pm 0.20$	$100.000 \pm 0.000$
-150 m < COGZ < 50 m	200	7	$15.83 \pm 0.84$	$100.000 \pm 0.000$
COGZ > 50 m	300	9	$17.67 \pm 1.23$	$100.000 \pm 0.000$

**Tabelle 2:** Die Tabelle zeigt die für den Gesamtprozess ausgewählten Parameter. Zugehörige Precision und Recall sind ebenfalls für jeden Cluster angegeben.

### 5.3 Ergebnis des Gesamtprozesses

Für den Gesamtprozess wurde ein Recall von

$$R = (50.74 \pm 1.86)\%$$

bei einer Precision von

$$P = (99.986 \pm 0.003)\%$$

erreicht. Diese Angaben beziehen sich auf Daten, auf welche die Precuts aus [3] bereits angewendet wurden. Die Fehler wurden mittels Fehlerfortpflanzung ermittelt.

Abschließend soll der Prozess noch einmal mit gewichteten Daten durchgeführt werden. Um mit anderen Analysen vergleichbar zu sein, werden die Daten auf einen Zeitraum von 33.281 Tagen normiert und auf ein  $E^{-3.7}$  Spektrum umgewichtet. Das geschieht indem für jedes Ereignis ein neues Gewicht nach den Neutrinoflussmodellen in [23] und [24] berechnet wird. Nach den in Abschnitt 5 erwähnten Precuts sind noch 10173 Neutrinos im Datensatz enthalten. Davon wurden 5390 mit einer Precision von  $100^{+0}_{-0.1}\%$  gefunden. Der reale Recall der Analyse beträgt also  $(53 \pm 2)\%$ . Die Fehler ergeben sich durch eine Fehlerfortpflanzung und wurden nach statistischen Überlegungen vergrößert. Die hohe Precision könnte an systematischen Fehlern liegen, die in Abschnitt 6 näher erläutert werden.

Analyse	Neutrinos	Precision
Diese Analyse	5390	$100^{+0}_{-0.1}\%$
Nach [4]	3909	$99.9^{+0.1}_{-0.2}\%$
Nach [4]	4815	$(98.7 \pm 0.4)\%$
Nach [3]	2833	$99.8^{+0.2}_{-0.4}\%$
Nach [3]	3638	$(99.4 \pm 0.6)\%$

**Tabelle 3:** Die Tabelle zeigt die Ergebnisse dieser Analyse im Vergleich zu Vorherigen. Alle Analysen verwenden die IC59 Geometrie, vgl. [4].

## 6 Resümee und Ausblick

Das Ziel dieser Arbeit war sowohl eine Maximierung des Recalls, als auch der Precision eines Multiensemblemodells zur Klassifizierung von IceCube-Monte-Carlo-Daten. Hierfür wurden im Rahmen dieser Arbeit neun Cluster auf vier verschiedenen Attributen definiert und einzeln optimiert. Dabei wurden geometrische, physikalische und statistische Überlegungen unternommen, um die Cluster sinnvoll auszuwählen. Zur Trennung der Daten wurde der Random Forest Algorithmus verwendet. Dieser wurde so eingestellt, dass er mit möglichst großer Präzision Aussagen über die Klasse eines Ereignisses liefert. Die Modelle, die auf den einzelnen Clustern gebildet wurden, lieferten durch ein Zusammenfügen der klassifizierten Ereignisse nach der in Abschnitt 5.2 beschriebenen Methode einen Recall von  $R = (50.74 \pm 1.86)\%$  bei einer Precision von  $P = (99.986 \pm 0.003)\%$ . Den höchsten Recall wies der Cluster  $RLogL < 9.5$  auf, was darauf schließen lässt, dass Ereignisse, bei denen die Rekonstruktion funktioniert hat, sich auch gut klassifizieren lassen.

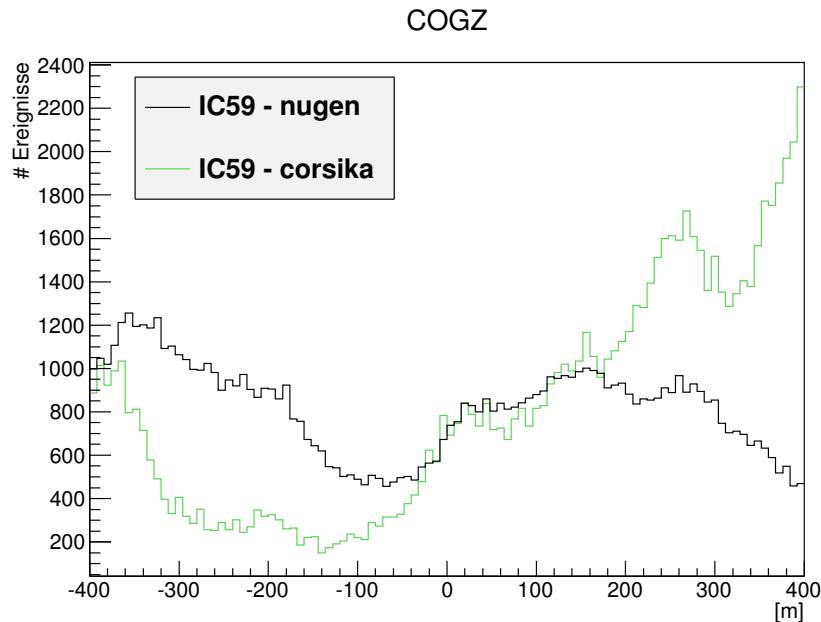
Die Analyse eines  $E^{-3.7}$  Spektrum ergab einen realen Recall von  $R = (53 \pm 2)\%$  bei einer Precision von  $P = 100^{+0}_{-0.1}\%$ . Ein Vergleich mit anderen Analysen findet sich in Tabelle 3. Es lässt sich eine Verbesserung des Recalls um ca. 10.6% bezüglich [4] bei ähnlicher Precision feststellen.

Alle angegebenen Fehler sind statistische Fehler. Es müssen aber zusätzlich systematische Fehler in Betracht gezogen werden. Da es sich bei den verwendeten Daten um Monte-Carlo-Daten handelt, kann es sein, dass der Random Forest Artefakte von Pseudo-Zufalls- oder Simulationsalgorithmen in den Daten findet. Auch das Ändern eines Zufalls-Seeds beim Trainieren der Multiensemblealgorithmen und der Vorverarbeitung der Examples kann zu Schwankungen in den Ergebnissen führen. Dieses Verhalten hat vermutlich zu unterschiedlichen Precisions in gewichteten und ungewichteten Daten

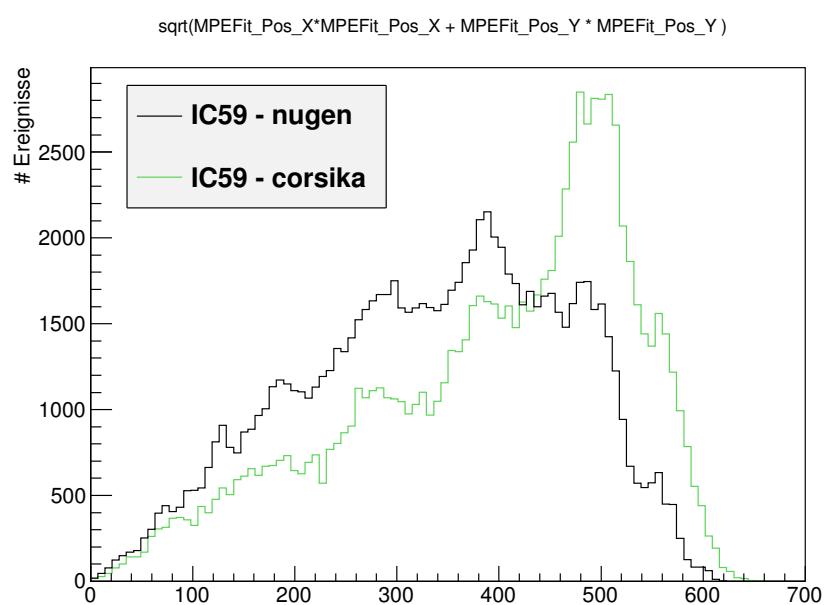
geführt. Hinzu kommt für eine richtige Analyse, dass es inhärente Unterschiede zwischen den Monte-Carlo-Daten und gemessenen Daten gibt. Es ist also nicht zwingend, dass die verwendeten Methoden auch auf Messdaten gute Resultate liefern. Die Größe der Statistik spielt ebenfalls eine Rolle. Aufgrund der starken Verunreinigung des Signals durch den Untergrund wäre es vermutlich sinnvoll den Random Forest mit einem größeren Hintergrund-zu-Signal-Verhältnis zu trainieren. Die Ergebnisse dieser und vorhergehender Arbeiten legen die Annahme nahe, dass es noch eine Vielzahl weiterer Cluster gibt, die höhere Erkennungsraten bei größerer Reinheit liefern. Clusteralgorithmen, wie der k-Means [25], können unabhängig von physikalischen Überlegungen Muster in Datenmengen suchen und damit den Separationsprozess automatisieren.

## 7 Anhang

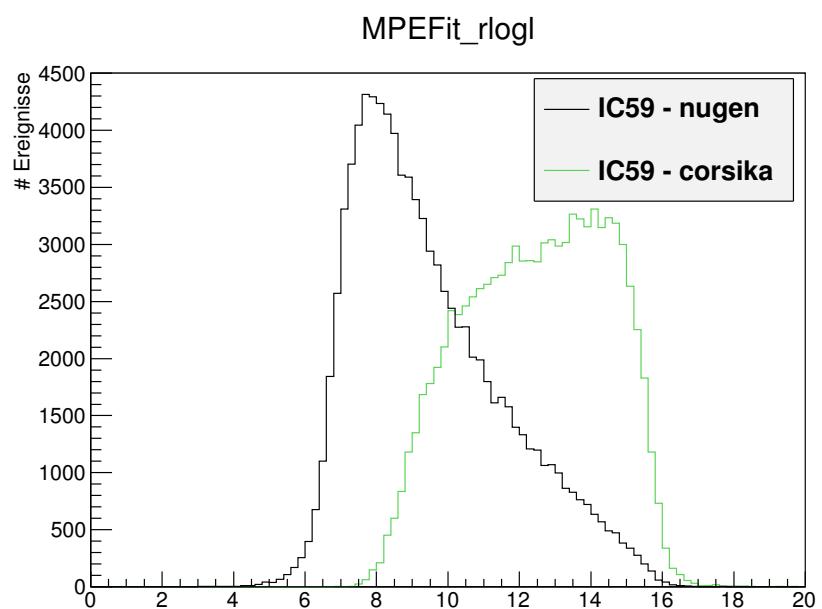
### Abbildungen



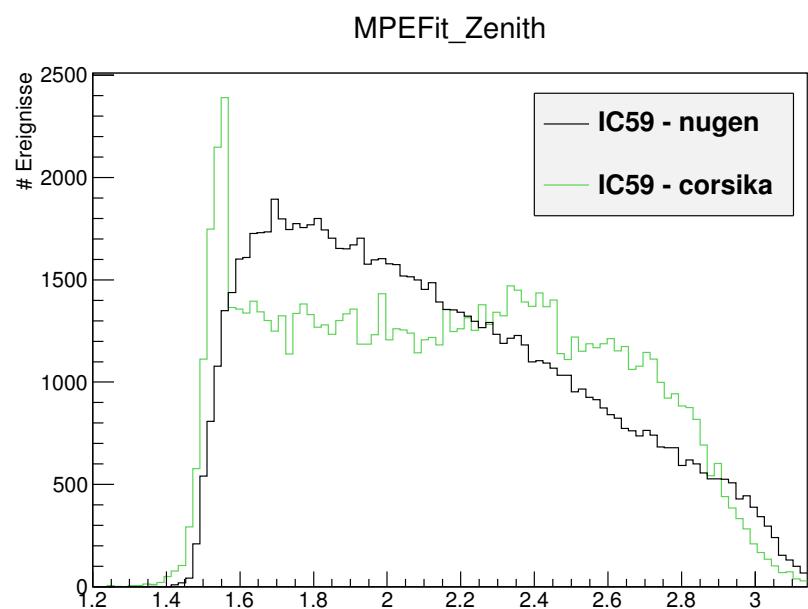
**Abbildung 7:** Das Histogramm für COGZ zeigt deutlich die Position des Dustlayers zwischen  $-150\text{ m}$  und  $50\text{ m}$ . An dieser Stelle gibt es einen deutlichen Einbruch im NuGen Spektrum. Ebenfalls gut zu erkennen ist, dass im oberen Bereich des Detektors mehr Hintergrundereignisse gemessen werden. (vgl. [4])



**Abbildung 8:** Darstellung der Verteilungen für den Abstand des Vertex von der Detektorhauptachse. Im äußeren Bereich des Detektors treten mehr Hintergrundereignisse auf.

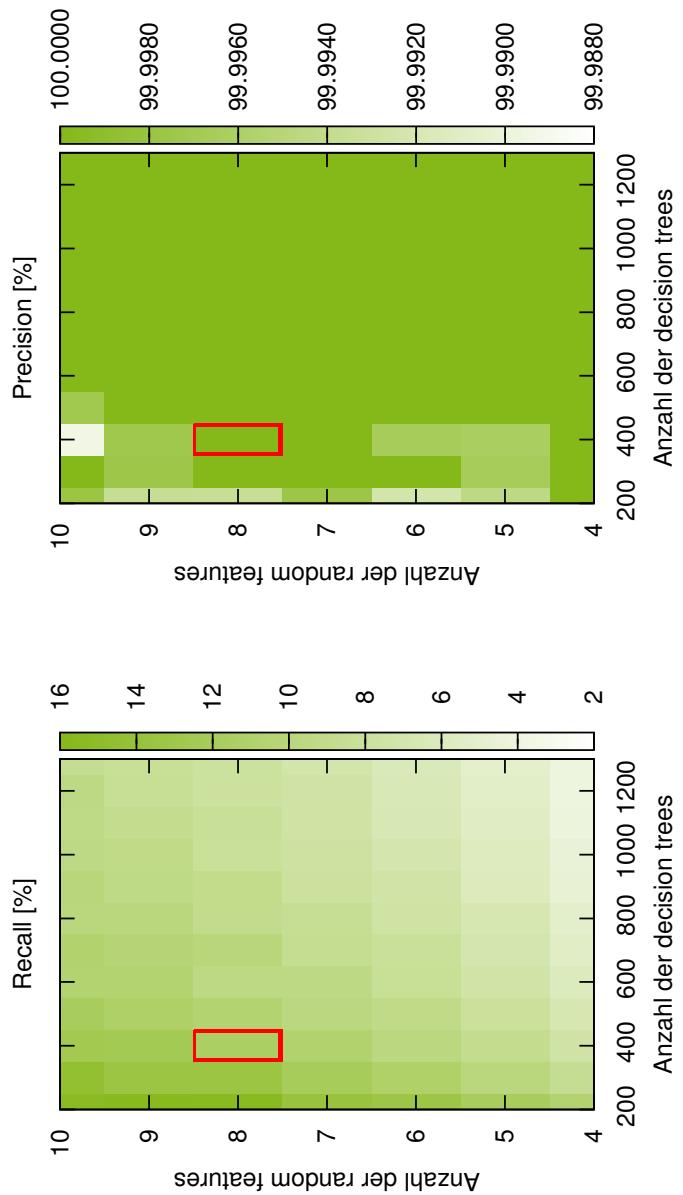


**Abbildung 9:** Vergleich zwischen den Verteilungen von Hintergrund und Signal für das Attribut RLogL.

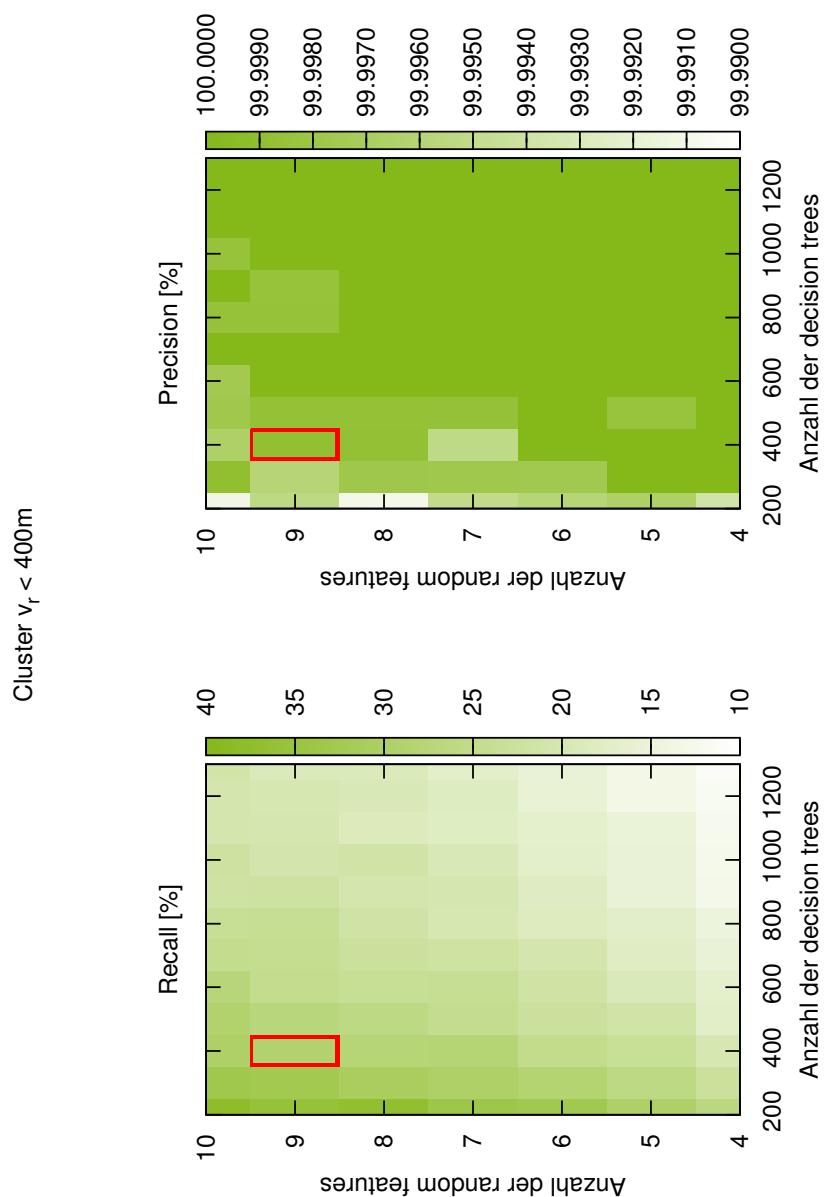


**Abbildung 10:** Darstellung der Verteilung des Zenitwinkels.

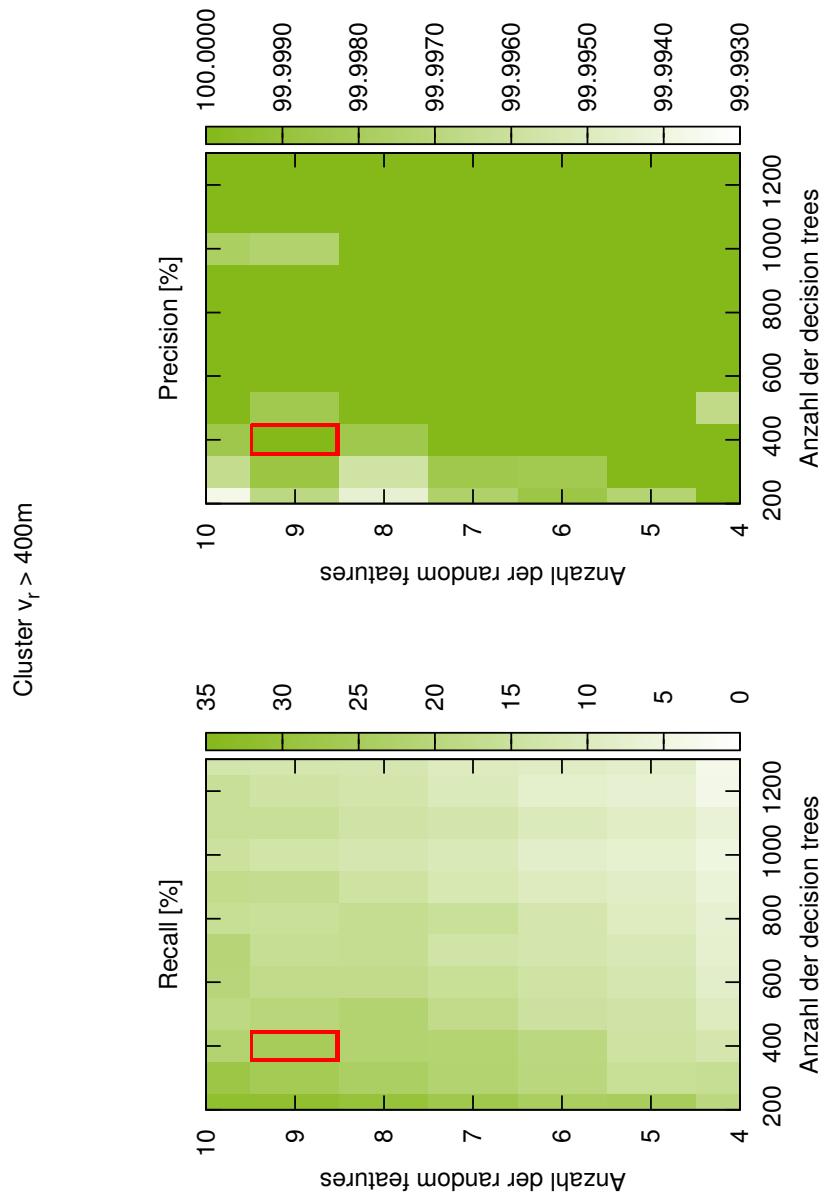
Cluster Zenit > 2.35 rad



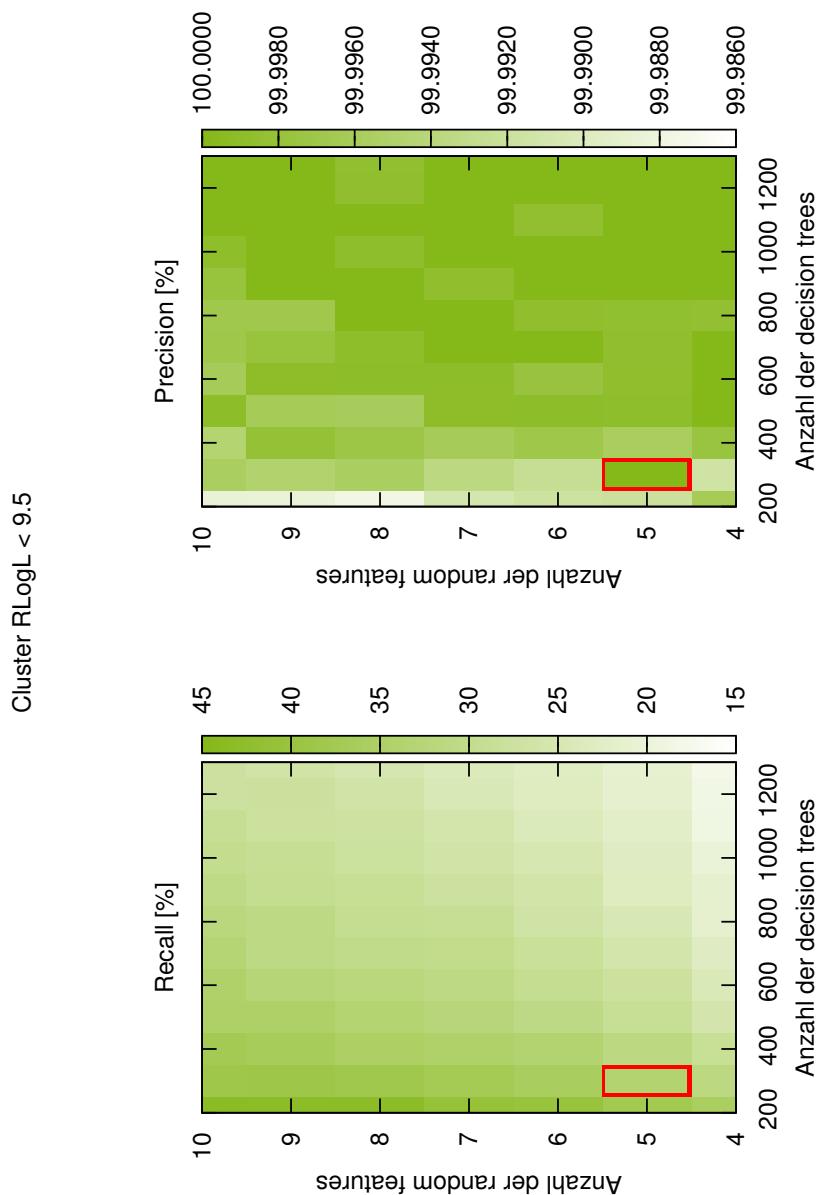
**Abbildung 11:** Der Plot zeigt Precision und Recall für den Cluster Zenitwinkel > 2.35. Der Plot ist farblich codiert. Je dunkler, desto größer sind Precision oder Recall. Das rote Rechteck zeigt, welche Einstellungen für den Gesamtprozess verwendet wurden.



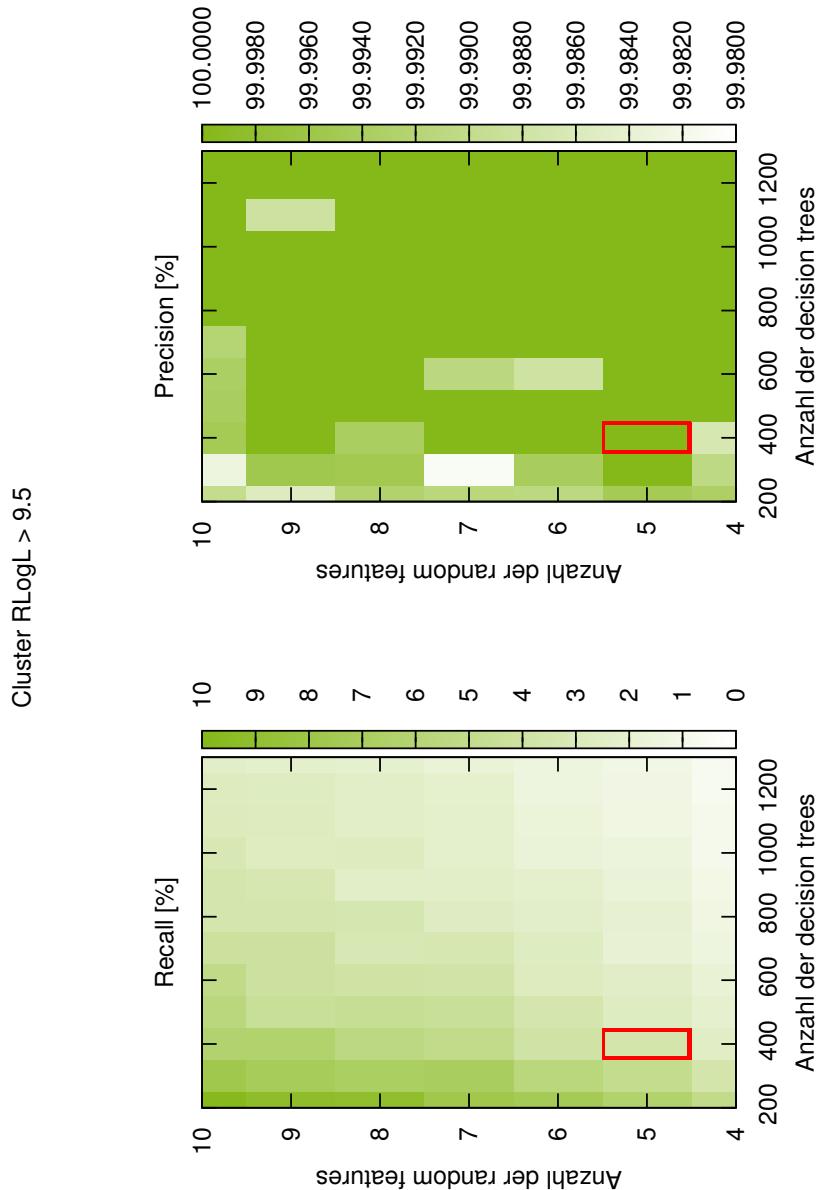
**Abbildung 12:** Der Plot zeigt Precision und Recall für den Cluster  $\text{vertex\_radius} < 400 \text{ m}$ .



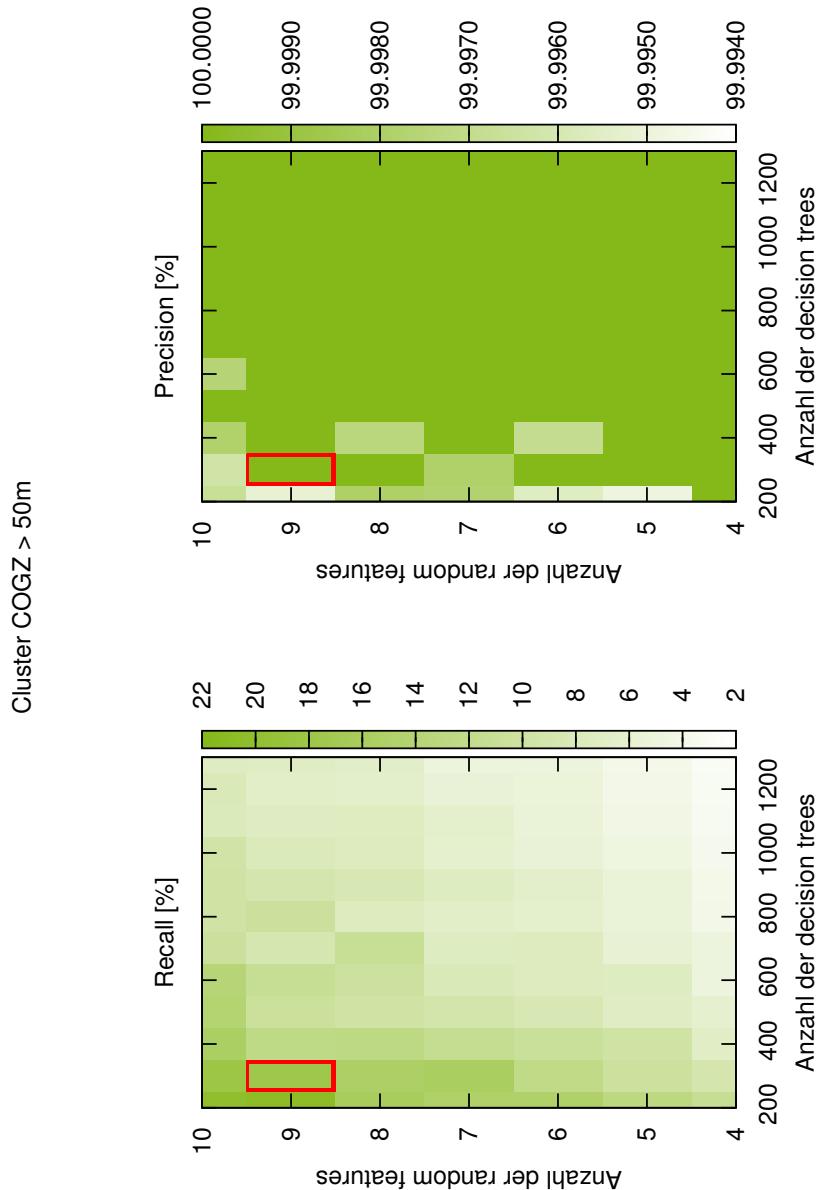
**Abbildung 13:** Der Plot zeigt Precision und Recall für den Cluster  $\text{vertex\_radius} > 400 \text{ m}$ .



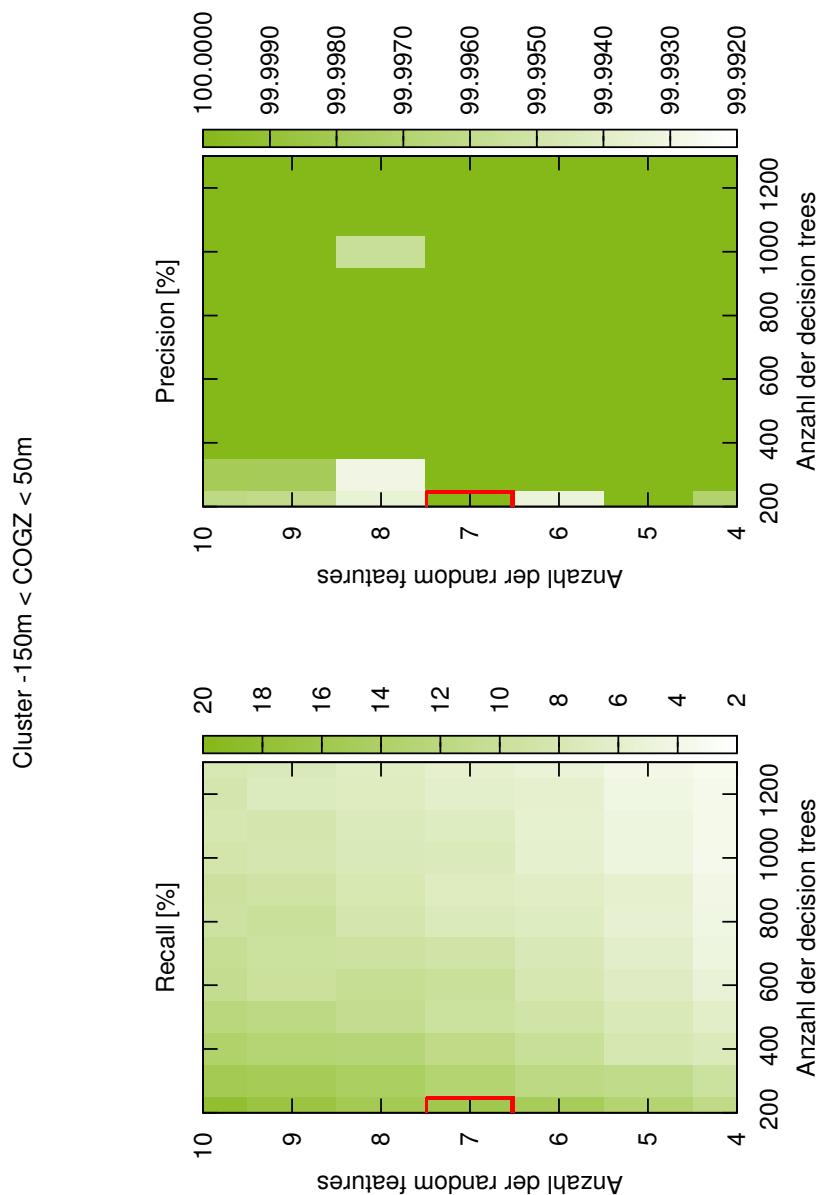
**Abbildung 14:** Der Plot zeigt Precision und Recall für den Cluster rlogl < 9.5.



**Abbildung 15:** Der Plot zeigt Precision und Recall für den Cluster rlogl > 9.5.



**Abbildung 16:** Der Plot zeigt Precision und Recall für den Cluster COGZ > 50 m.



**Abbildung 17:** Der Plot zeigt Precision und Recall für den Cluster  $-150 \text{ m} < \text{COGZ} < 50 \text{ m}$ .

Cluster COGZ < -150m

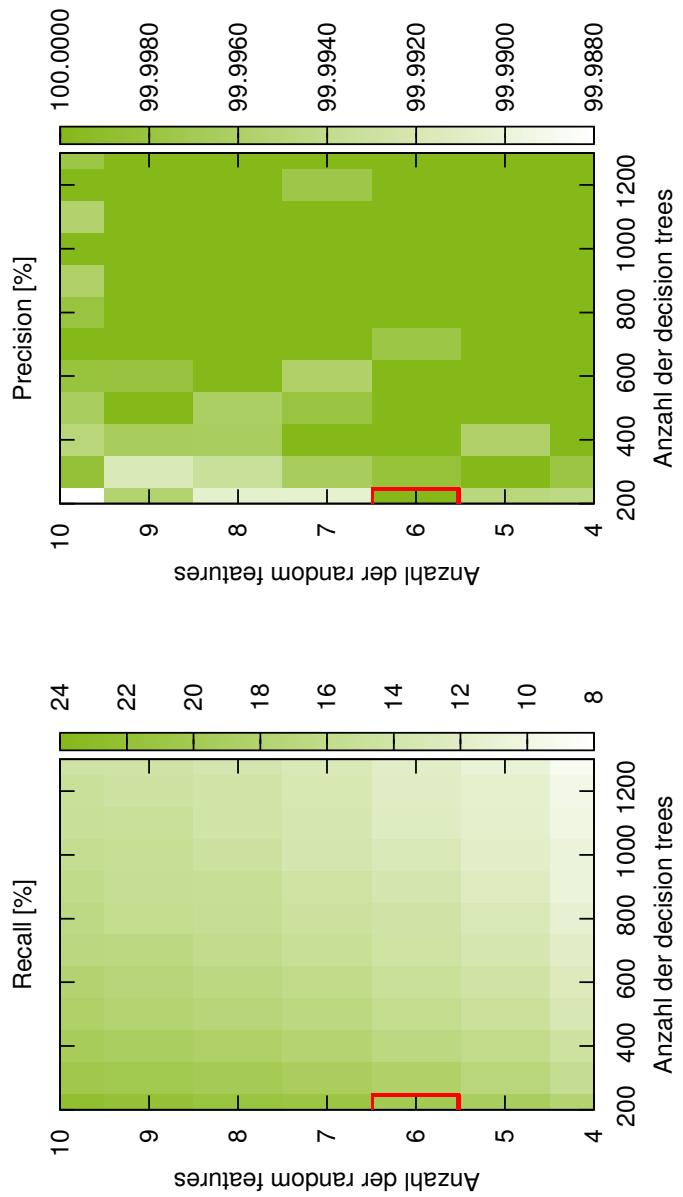


Abbildung 18: Der Plot zeigt Precision und Recall für den Cluster COGZ < -150 m.

## Tabellen

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
200	4	26.71	0.37	99.994	0.000
200	5	30.00	1.46	99.997	0.001
200	6	32.96	1.27	99.996	0.001
200	7	33.69	0.24	99.995	0.002
200	8	36.92	0.12	99.991	0.002
200	9	35.83	0.83	99.995	0.000
200	10	37.80	0.31	99.991	0.004
300	4	23.03	0.44	100.000	0.000
300	5	25.97	0.50	100.000	0.000
300	6	28.40	0.23	99.998	0.001
300	7	30.02	0.71	99.998	0.001
300	8	30.87	0.22	99.998	0.001
300	9	32.71	0.54	99.996	0.003
300	10	33.40	0.28	99.999	0.000
400	4	20.17	0.47	100.000	0.000
400	5	23.65	0.64	100.000	0.000
400	6	24.97	0.26	100.000	0.000
400	7	28.15	0.71	99.995	0.002
400	8	27.88	0.83	99.999	0.001
400	9	28.83	0.59	99.999	0.001
400	10	29.81	0.38	99.997	0.002
500	4	17.22	0.14	100.000	0.000
500	5	21.50	0.27	99.998	0.001
500	6	22.95	0.61	100.000	0.000
500	7	24.56	0.10	99.999	0.001
500	8	26.13	0.57	99.999	0.001
500	9	27.38	0.47	99.999	0.002
500	10	29.13	0.04	99.998	0.002
600	4	16.78	0.29	100.000	0.000
600	5	19.16	0.05	100.000	0.000
600	6	21.77	0.36	100.000	0.000
600	7	24.00	0.17	100.000	0.000

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
600	8	23.90	0.21	100.000	0.000
600	9	24.68	0.27	100.000	0.000
600	10	27.46	0.66	99.998	0.001
700	4	15.29	0.39	100.000	0.000
700	5	17.72	0.14	100.000	0.000
700	6	20.61	0.35	100.000	0.000
700	7	22.08	0.09	100.000	0.000
700	8	22.99	0.81	100.000	0.000
700	9	24.45	0.32	100.000	0.000
700	10	24.75	0.47	100.000	0.000
800	4	14.61	0.09	100.000	0.000
800	5	17.11	0.27	100.000	0.000
800	6	18.53	0.67	100.000	0.000
800	7	20.19	0.58	100.000	0.000
800	8	21.65	0.51	100.000	0.000
800	9	23.99	0.52	99.999	0.001
800	10	23.92	0.00	99.999	0.003
900	4	12.97	0.38	100.000	0.000
900	5	15.48	0.36	100.000	0.000
900	6	17.95	0.19	100.000	0.000
900	7	20.26	0.02	100.000	0.000
900	8	20.74	0.59	100.000	0.000
900	9	22.25	0.71	99.999	0.001
900	10	22.04	0.89	100.000	0.000
1000	4	12.62	0.31	100.000	0.000
1000	5	15.34	0.18	100.000	0.000
1000	6	16.82	0.69	100.000	0.000
1000	7	19.56	0.03	100.000	0.000
1000	8	21.30	0.70	100.000	0.000
1000	9	20.84	0.50	100.000	0.000
1000	10	22.33	0.48	99.999	0.001
1100	4	12.26	0.09	100.000	0.000
1100	5	15.06	0.76	100.000	0.000
1100	6	16.62	0.42	100.000	0.000

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
1100	7	17.74	0.25	100.000	0.000
1100	8	18.68	0.47	100.000	0.000
1100	9	20.24	0.24	100.000	0.000
1100	10	20.69	0.62	100.000	0.000
1200	4	11.65	0.09	100.000	0.000
1200	5	13.30	0.64	100.000	0.000
1200	6	15.53	0.59	100.000	0.000
1200	7	18.16	0.12	100.000	0.000
1200	8	19.57	0.69	100.000	0.000
1200	9	19.97	0.52	100.000	0.000
1200	10	20.63	0.44	100.000	0.000
1300	4	11.07	0.22	100.000	0.000
1300	5	13.33	0.34	100.000	0.000
1300	6	15.50	0.07	100.000	0.000
1300	7	16.96	0.72	100.000	0.000
1300	8	18.83	0.12	100.000	0.000
1300	9	19.09	0.22	100.000	0.000
1300	10	21.19	0.02	100.000	0.000

**Tabelle 4:** Precision und Recall, sowie deren Fehler, für den Cluster  $v_r < 400m$  in Abhängigkeit von *I* und *K*.

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
200	4	19.48	1.54	100.000	0.000
200	5	24.82	0.78	99.997	0.002
200	6	24.14	0.29	99.999	0.001
200	7	27.24	1.04	99.998	0.002
200	8	29.80	1.02	99.994	0.000
200	9	31.37	0.04	99.997	0.001
200	10	31.58	0.42	99.994	0.002
300	4	16.34	1.07	100.000	0.000
300	5	15.98	0.68	100.000	0.000
300	6	19.75	0.58	99.998	0.003
300	7	21.65	0.39	99.998	0.003
300	8	23.73	1.52	99.996	0.002
300	9	25.82	1.26	99.999	0.001
300	10	27.99	1.17	99.996	0.002
400	4	12.43	0.05	100.000	0.000
400	5	14.17	0.41	100.000	0.000
400	6	19.46	1.44	100.000	0.000
400	7	21.41	0.52	100.000	0.000
400	8	21.75	0.37	99.998	0.003
400	9	25.09	0.09	100.000	0.000
400	10	21.82	1.04	99.998	0.001
500	4	9.89	1.21	99.997	0.002
500	5	13.66	1.88	100.000	0.000
500	6	14.61	2.41	100.000	0.000
500	7	17.81	1.69	100.000	0.000
500	8	21.73	0.72	100.000	0.000
500	9	20.20	2.47	99.998	0.003
500	10	18.86	3.81	100.000	0.000
600	4	8.32	0.84	100.000	0.000
600	5	11.95	0.58	100.000	0.000
600	6	13.72	0.75	100.000	0.000
600	7	15.86	0.15	100.000	0.000
600	8	17.76	1.49	100.000	0.000
600	9	17.79	2.42	100.000	0.000

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
600	10	20.41	0.20	100.000	0.000
700	4	7.36	0.83	100.000	0.000
700	5	11.20	1.47	100.000	0.000
700	6	12.49	2.29	100.000	0.000
700	7	13.51	0.77	100.000	0.000
700	8	16.75	1.89	100.000	0.000
700	9	16.37	1.60	100.000	0.000
700	10	20.71	0.34	100.000	0.000
800	4	6.85	0.12	100.000	0.000
800	5	9.60	1.53	100.000	0.000
800	6	12.35	1.71	100.000	0.000
800	7	15.50	0.44	100.000	0.000
800	8	16.90	0.15	100.000	0.000
800	9	15.54	1.92	100.000	0.000
800	10	16.10	3.25	100.000	0.000
900	4	5.93	0.75	100.000	0.000
900	5	8.69	0.30	100.000	0.000
900	6	9.88	1.25	100.000	0.000
900	7	11.58	1.53	100.000	0.000
900	8	14.09	1.56	100.000	0.000
900	9	17.06	0.60	100.000	0.000
900	10	17.45	0.06	100.000	0.000
1000	4	4.65	0.95	100.000	0.000
1000	5	7.16	1.67	100.000	0.000
1000	6	8.31	1.17	100.000	0.000
1000	7	11.06	1.96	100.000	0.000
1000	8	12.16	1.56	100.000	0.000
1000	9	13.01	2.07	99.997	0.005
1000	10	14.94	1.52	99.998	0.001
1100	4	6.19	0.80	100.000	0.000
1100	5	8.86	0.88	100.000	0.000
1100	6	10.42	1.27	100.000	0.000
1100	7	12.70	1.73	100.000	0.000
1100	8	13.63	1.60	100.000	0.000

$I$	$K$	$R[\%]$	$\delta R[\%]$	$P[\%]$	$\delta P[\%]$
1100	9	15.69	0.71	100.000	0.000
1100	10	15.85	1.56	100.000	0.000
1200	4	3.76	0.58	100.000	0.000
1200	5	7.10	0.87	100.000	0.000
1200	6	7.66	1.35	100.000	0.000
1200	7	10.60	1.96	100.000	0.000
1200	8	12.73	2.41	100.000	0.000
1200	9	13.77	0.02	100.000	0.000
1200	10	15.69	0.81	100.000	0.000
1300	4	3.56	0.39	100.000	0.000
1300	5	8.00	0.93	100.000	0.000
1300	6	8.97	0.27	100.000	0.000
1300	7	9.94	0.34	100.000	0.000
1300	8	12.14	0.83	100.000	0.000
1300	9	12.35	0.76	100.000	0.000
1300	10	12.86	1.46	100.000	0.000

**Tabelle 5:** Precision und Recall, sowie deren Fehler, für den Cluster  $v_r > 400m$  in Abhängigkeit von  $I$  und  $K$ .

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
200	4	11.32	0.78	100.000	0.000
200	5	12.97	0.33	99.995	0.004
200	6	15.03	1.07	99.996	0.002
200	7	14.93	1.23	99.998	0.001
200	8	16.22	0.93	99.998	0.001
200	9	20.71	1.10	99.995	0.002
200	10	20.27	0.13	99.997	0.007
300	4	9.09	0.58	100.000	0.000
300	5	10.59	0.58	100.000	0.000
300	6	12.33	1.05	100.000	0.000
300	7	15.84	1.29	99.998	0.004
300	8	15.43	0.01	100.000	0.000
300	9	17.67	1.23	100.000	0.000
300	10	18.05	2.09	99.996	0.003
400	4	7.07	0.72	100.000	0.000
400	5	9.90	1.39	100.000	0.000
400	6	10.82	0.51	99.997	0.002
400	7	11.55	0.94	100.000	0.000
400	8	12.67	0.94	99.997	0.001
400	9	12.51	1.06	100.000	0.000
400	10	15.53	3.41	99.998	0.001
500	4	6.26	0.07	100.000	0.000
500	5	7.09	0.01	100.000	0.000
500	6	8.49	0.14	100.000	0.000
500	7	9.38	0.08	100.000	0.000
500	8	9.85	0.09	100.000	0.000
500	9	10.73	0.52	100.000	0.000
500	10	14.33	0.81	100.000	0.000
600	4	5.03	0.06	100.000	0.000
600	5	7.42	0.22	100.000	0.000
600	6	7.62	0.08	100.000	0.000
600	7	8.18	0.01	100.000	0.000
600	8	10.35	0.64	100.000	0.000
600	9	11.37	0.09	100.000	0.000

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
600	10	13.82	0.86	99.998	0.001
700	4	5.02	0.10	100.000	0.000
700	5	5.86	0.07	100.000	0.000
700	6	7.63	0.28	100.000	0.000
700	7	7.44	0.12	100.000	0.000
700	8	11.18	1.01	100.000	0.000
700	9	8.87	0.01	100.000	0.000
700	10	10.64	0.84	100.000	0.000
800	4	4.16	0.02	100.000	0.000
800	5	5.51	0.44	100.000	0.000
800	6	6.42	0.04	100.000	0.000
800	7	6.91	0.32	100.000	0.000
800	8	7.70	0.31	100.000	0.000
800	9	10.63	0.59	100.000	0.000
800	10	9.86	0.36	100.000	0.000
900	4	3.94	0.10	100.000	0.000
900	5	5.43	0.38	100.000	0.000
900	6	6.58	0.03	100.000	0.000
900	7	7.36	0.04	100.000	0.000
900	8	8.45	0.59	100.000	0.000
900	9	9.10	0.10	100.000	0.000
900	10	9.93	0.69	100.000	0.000
1000	4	3.61	0.06	100.000	0.000
1000	5	4.71	0.17	100.000	0.000
1000	6	5.63	0.48	100.000	0.000
1000	7	6.25	0.14	100.000	0.000
1000	8	7.65	0.07	100.000	0.000
1000	9	8.04	0.22	100.000	0.000
1000	10	9.64	0.75	100.000	0.000
1100	4	3.35	0.15	100.000	0.000
1100	5	4.41	0.18	100.000	0.000
1100	6	5.51	0.01	100.000	0.000
1100	7	6.53	0.65	100.000	0.000
1100	8	7.50	0.14	100.000	0.000

$I$	$K$	$R[\%]$	$\delta R[\%]$	$P[\%]$	$\delta P[\%]$
1100	9	7.29	0.17	100.000	0.000
1100	10	7.87	0.01	100.000	0.000
1200	4	3.16	0.07	100.000	0.000
1200	5	4.12	0.05	100.000	0.000
1200	6	5.34	0.00	100.000	0.000
1200	7	5.60	0.14	100.000	0.000
1200	8	6.63	0.08	100.000	0.000
1200	9	6.84	0.63	100.000	0.000
1200	10	8.23	0.03	100.000	0.000
1300	4	2.98	0.02	100.000	0.000
1300	5	4.10	0.02	100.000	0.000
1300	6	5.04	0.57	100.000	0.000
1300	7	5.19	0.32	100.000	0.000
1300	8	6.74	0.73	100.000	0.000
1300	9	7.09	0.14	100.000	0.000
1300	10	7.29	0.26	100.000	0.000

**Tabelle 6:** Precision und Recall, sowie deren Fehler, für den Cluster  $COGZ > 50m$  in Abhängigkeit von  $I$  und  $K$ .

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
200	4	11.34	0.63	99.997	0.001
200	5	12.99	0.18	100.000	0.000
200	6	14.90	0.49	99.993	0.008
200	7	15.83	0.84	100.000	0.000
200	8	15.42	0.68	99.994	0.003
200	9	16.87	0.35	99.996	0.003
200	10	18.13	0.49	99.996	0.003
300	4	9.64	0.24	100.000	0.000
300	5	11.23	0.32	100.000	0.000
300	6	11.62	1.07	100.000	0.000
300	7	13.07	0.56	100.000	0.000
300	8	14.25	1.00	99.993	0.002
300	9	15.09	0.44	99.998	0.001
300	10	15.49	0.65	99.998	0.001
400	4	7.42	0.02	100.000	0.000
400	5	8.00	0.28	100.000	0.000
400	6	10.07	0.44	100.000	0.000
400	7	11.34	0.88	100.000	0.000
400	8	12.70	0.74	100.000	0.000
400	9	12.91	1.11	100.000	0.000
400	10	13.57	0.50	100.000	0.000
500	4	6.29	0.56	100.000	0.000
500	5	7.66	0.44	100.000	0.000
500	6	8.81	0.31	100.000	0.000
500	7	9.64	0.08	100.000	0.000
500	8	10.76	0.62	100.000	0.000
500	9	11.75	0.49	100.000	0.000
500	10	12.20	0.36	100.000	0.000
600	4	5.22	0.55	100.000	0.000
600	5	6.74	0.17	100.000	0.000
600	6	8.05	1.28	100.000	0.000
600	7	9.89	0.74	100.000	0.000
600	8	10.41	0.53	100.000	0.000
600	9	9.87	1.00	100.000	0.000

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
600	10	10.76	0.31	100.000	0.000
700	4	4.60	0.21	100.000	0.000
700	5	6.22	0.37	100.000	0.000
700	6	7.75	0.13	100.000	0.000
700	7	8.75	0.16	100.000	0.000
700	8	9.40	0.25	100.000	0.000
700	9	9.61	0.89	100.000	0.000
700	10	10.47	0.57	100.000	0.000
800	4	4.34	0.00	100.000	0.000
800	5	5.55	0.17	100.000	0.000
800	6	6.86	0.75	100.000	0.000
800	7	7.31	0.07	100.000	0.000
800	8	8.42	0.11	100.000	0.000
800	9	9.88	0.04	100.000	0.000
800	10	9.31	0.04	100.000	0.000
900	4	4.12	0.05	100.000	0.000
900	5	5.78	0.15	100.000	0.000
900	6	6.57	0.28	100.000	0.000
900	7	6.97	0.03	100.000	0.000
900	8	7.94	0.36	100.000	0.000
900	9	9.05	0.00	100.000	0.000
900	10	9.53	0.96	100.000	0.000
1000	4	3.57	0.16	100.000	0.000
1000	5	4.58	0.14	100.000	0.000
1000	6	5.87	0.20	100.000	0.000
1000	7	7.29	0.93	100.000	0.000
1000	8	7.53	0.02	99.996	0.002
1000	9	8.13	1.18	100.000	0.000
1000	10	8.56	0.51	100.000	0.000
1100	4	3.66	0.21	100.000	0.000
1100	5	4.55	0.13	100.000	0.000
1100	6	5.66	0.30	100.000	0.000
1100	7	6.86	0.63	100.000	0.000
1100	8	7.33	0.26	100.000	0.000

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
1100	9	8.35	0.32	100.000	0.000
1100	10	8.01	0.94	100.000	0.000
1200	4	3.61	0.03	100.000	0.000
1200	5	4.36	0.26	100.000	0.000
1200	6	5.79	0.44	100.000	0.000
1200	7	6.18	0.46	100.000	0.000
1200	8	6.94	0.29	100.000	0.000
1200	9	7.14	0.13	100.000	0.000
1200	10	8.45	1.01	100.000	0.000
1300	4	3.31	0.02	100.000	0.000
1300	5	4.03	0.09	100.000	0.000
1300	6	5.07	0.22	100.000	0.000
1300	7	5.74	0.44	100.000	0.000
1300	8	6.71	0.12	100.000	0.000
1300	9	7.51	1.17	100.000	0.000
1300	10	7.77	0.89	100.000	0.000

**Tabelle 7:** Precision und Recall, sowie deren Fehler, für den Cluster  $-150m < COGZ < 50m$  in Abhängigkeit von *I* und *K*.

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
200	4	17.80	0.17	99.994	0.002
200	5	19.20	0.07	99.995	0.003
200	6	20.03	0.20	100.000	0.000
200	7	20.94	0.18	99.990	0.005
200	8	21.31	0.07	99.991	0.005
200	9	21.89	0.11	99.995	0.002
200	10	22.32	0.10	99.988	0.006
300	4	15.57	0.02	99.998	0.001
300	5	17.16	0.12	100.000	0.000
300	6	18.47	0.11	99.998	0.001
300	7	19.17	0.03	99.997	0.007
300	8	19.84	0.28	99.993	0.003
300	9	20.17	0.07	99.992	0.004
300	10	20.32	0.22	99.998	0.003
400	4	14.59	0.05	100.000	0.000
400	5	15.97	0.07	99.996	0.003
400	6	16.82	0.12	100.000	0.000
400	7	17.84	0.05	100.000	0.000
400	8	18.57	0.41	99.996	0.002
400	9	19.04	0.02	99.996	0.002
400	10	19.49	0.03	99.995	0.006
500	4	13.19	0.26	100.000	0.000
500	5	14.97	0.08	100.000	0.000
500	6	15.88	0.05	100.000	0.000
500	7	16.87	0.26	99.998	0.001
500	8	17.33	0.05	99.996	0.003
500	9	17.76	0.04	100.000	0.000
500	10	18.54	0.15	99.996	0.002
600	4	12.67	0.03	100.000	0.000
600	5	14.31	0.10	100.000	0.000
600	6	15.12	0.19	100.000	0.000
600	7	16.10	0.12	99.996	0.002
600	8	16.86	0.15	100.000	0.000
600	9	17.32	0.01	99.998	0.001

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
600	10	17.92	0.10	99.998	0.004
700	4	11.95	0.13	100.000	0.000
700	5	13.54	0.05	100.000	0.000
700	6	14.49	0.08	99.998	0.001
700	7	15.29	0.07	100.000	0.000
700	8	16.03	0.19	100.000	0.000
700	9	16.85	0.01	100.000	0.000
700	10	16.91	0.26	100.000	0.000
800	4	11.22	0.04	100.000	0.000
800	5	13.07	0.14	100.000	0.000
800	6	14.20	0.12	100.000	0.000
800	7	14.84	0.16	100.000	0.000
800	8	15.45	0.01	100.000	0.000
800	9	15.93	0.28	100.000	0.000
800	10	16.56	0.09	99.998	0.001
900	4	10.58	0.15	100.000	0.000
900	5	12.41	0.13	100.000	0.000
900	6	13.62	0.13	100.000	0.000
900	7	14.48	0.07	100.000	0.000
900	8	15.33	0.14	100.000	0.000
900	9	15.49	0.05	100.000	0.000
900	10	16.19	0.02	99.996	0.002
1000	4	10.58	0.12	100.000	0.000
1000	5	11.86	0.04	100.000	0.000
1000	6	13.08	0.05	100.000	0.000
1000	7	13.59	0.09	100.000	0.000
1000	8	14.63	0.18	100.000	0.000
1000	9	15.34	0.05	100.000	0.000
1000	10	15.65	0.06	100.000	0.000
1100	4	10.02	0.15	100.000	0.000
1100	5	11.57	0.06	100.000	0.000
1100	6	12.30	0.11	100.000	0.000
1100	7	13.44	0.06	100.000	0.000
1100	8	14.02	0.33	100.000	0.000

$I$	$K$	$R[\%]$	$\delta R[\%]$	$P[\%]$	$\delta P[\%]$
1100	9	15.04	0.07	100.000	0.000
1100	10	15.20	0.10	99.996	0.002
1200	4	9.69	0.19	100.000	0.000
1200	5	11.48	0.05	100.000	0.000
1200	6	12.11	0.23	100.000	0.000
1200	7	13.26	0.24	99.997	0.001
1200	8	14.02	0.03	100.000	0.000
1200	9	14.61	0.08	100.000	0.000
1200	10	15.03	0.05	100.000	0.000
1300	4	9.15	0.12	100.000	0.000
1300	5	10.88	0.04	100.000	0.000
1300	6	11.99	0.05	100.000	0.000
1300	7	13.04	0.29	100.000	0.000
1300	8	13.60	0.11	100.000	0.000
1300	9	14.30	0.28	100.000	0.000
1300	10	14.38	0.33	99.998	0.001

**Tabelle 8:** Precision und Recall, sowie deren Fehler, für den Cluster  $COGZ < -150m$  in Abhängigkeit von  $I$  und  $K$ .

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
200	4	35.73	0.40	99.996	0.002
200	5	37.83	0.11	99.992	0.000
200	6	40.42	0.06	99.992	0.004
200	7	40.81	0.01	99.991	0.002
200	8	43.20	0.49	99.988	0.006
200	9	43.20	0.21	99.988	0.000
200	10	43.36	0.26	99.988	0.006
300	4	31.48	0.33	99.992	0.001
300	5	33.52	0.47	100.000	0.000
300	6	36.02	0.53	99.993	0.001
300	7	37.08	0.13	99.994	0.001
300	8	38.24	1.24	99.996	0.002
300	9	38.81	0.15	99.995	0.002
300	10	38.67	0.95	99.996	0.004
400	4	28.64	0.08	99.998	0.001
400	5	31.48	0.31	99.996	0.001
400	6	33.65	0.43	99.997	0.001
400	7	34.50	0.16	99.996	0.002
400	8	35.23	0.48	99.997	0.001
400	9	36.49	0.82	99.998	0.001
400	10	37.28	0.33	99.995	0.003
500	4	25.87	0.10	100.000	0.000
500	5	28.73	0.12	99.999	0.001
500	6	30.91	0.34	99.999	0.001
500	7	32.52	0.22	99.999	0.001
500	8	33.35	0.35	99.996	0.002
500	9	34.92	0.28	99.996	0.002
500	10	34.79	0.13	99.999	0.000
600	4	24.46	0.07	100.000	0.000
600	5	27.56	0.57	99.999	0.001
600	6	29.51	0.02	99.998	0.002
600	7	31.00	0.57	99.999	0.001
600	8	32.02	0.24	99.999	0.001
600	9	32.90	0.88	99.999	0.001

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
600	10	34.38	0.33	99.996	0.002
700	4	22.96	0.12	100.000	0.000
700	5	26.10	0.22	99.999	0.001
700	6	28.26	0.41	100.000	0.000
700	7	29.91	0.20	100.000	0.000
700	8	30.38	0.21	99.999	0.001
700	9	30.94	0.18	99.998	0.001
700	10	33.25	0.04	99.997	0.001
800	4	21.23	0.09	99.998	0.001
800	5	24.71	0.17	99.999	0.001
800	6	26.65	0.22	99.999	0.003
800	7	29.13	0.07	100.000	0.000
800	8	29.34	0.91	100.000	0.000
800	9	31.03	0.28	99.997	0.002
800	10	31.81	0.43	99.997	0.001
900	4	21.42	0.28	100.000	0.000
900	5	23.27	0.06	100.000	0.000
900	6	26.17	0.32	100.000	0.000
900	7	27.43	0.04	99.999	0.001
900	8	28.90	0.37	100.000	0.000
900	9	29.30	0.14	100.000	0.000
900	10	30.73	0.42	99.998	0.001
1000	4	20.27	0.28	100.000	0.000
1000	5	22.94	0.12	100.000	0.000
1000	6	25.06	0.21	100.000	0.000
1000	7	26.27	0.18	100.000	0.000
1000	8	27.71	0.64	99.999	0.001
1000	9	28.95	0.05	100.000	0.000
1000	10	29.71	0.20	99.999	0.001
1100	4	18.94	0.16	100.000	0.000
1100	5	22.22	0.10	100.000	0.000
1100	6	23.83	0.55	99.999	0.001
1100	7	26.03	0.02	100.000	0.000
1100	8	27.27	0.16	100.000	0.000

$I$	$K$	$R[\%]$	$\delta R[\%]$	$P[\%]$	$\delta P[\%]$
1100	9	27.60	0.36	100.000	0.000
1100	10	29.01	0.54	100.000	0.000
1200	4	18.43	0.08	100.000	0.000
1200	5	21.24	0.37	100.000	0.000
1200	6	23.21	0.17	100.000	0.000
1200	7	24.51	0.53	100.000	0.000
1200	8	26.43	0.21	99.999	0.003
1200	9	27.94	0.46	100.000	0.000
1200	10	27.54	0.01	100.000	0.000
1300	4	17.97	0.41	100.000	0.000
1300	5	20.91	0.54	100.000	0.000
1300	6	22.74	0.06	100.000	0.000
1300	7	23.97	0.07	100.000	0.000
1300	8	25.38	0.35	99.999	0.001
1300	9	26.38	0.11	100.000	0.000
1300	10	27.52	0.16	100.000	0.000

**Tabelle 9:** Precision und Recall, sowie deren Fehler, für den Cluster  $RLogL < 9.5$  in Abhängigkeit von  $I$  und  $K$ .

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
200	4	5.09	0.10	99.993	0.003
200	5	6.34	0.04	99.995	0.003
200	6	7.41	0.19	99.991	0.007
200	7	7.70	0.15	99.991	0.017
200	8	9.13	0.58	99.993	0.005
200	9	9.26	0.03	99.986	0.002
200	10	9.82	0.54	99.990	0.005
300	4	3.63	0.08	99.991	0.005
300	5	4.93	0.30	100.000	0.000
300	6	5.69	0.07	99.994	0.012
300	7	7.01	0.38	99.981	0.026
300	8	6.83	0.59	99.995	0.010
300	9	7.21	0.10	99.995	0.002
300	10	7.76	0.27	99.983	0.013
400	4	2.51	0.18	99.987	0.007
400	5	3.64	0.14	100.000	0.000
400	6	3.89	0.08	100.000	0.000
400	7	5.09	0.21	100.000	0.000
400	8	5.44	0.16	99.994	0.003
400	9	6.35	0.11	100.000	0.000
400	10	6.28	0.37	99.995	0.003
500	4	2.12	0.12	100.000	0.000
500	5	2.73	0.04	100.000	0.000
500	6	3.54	0.01	100.000	0.000
500	7	4.52	0.01	100.000	0.000
500	8	4.71	0.16	100.000	0.000
500	9	4.50	0.23	100.000	0.000
500	10	5.61	0.09	99.994	0.003
600	4	1.84	0.02	100.000	0.000
600	5	2.49	0.15	100.000	0.000
600	6	2.79	0.07	99.988	0.006
600	7	3.77	0.07	99.991	0.004
600	8	3.97	0.24	100.000	0.000
600	9	4.15	0.40	100.000	0.000

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
600	10	5.17	0.63	99.994	0.003
700	4	1.37	0.05	100.000	0.000
700	5	1.93	0.15	100.000	0.000
700	6	2.67	0.16	100.000	0.000
700	7	3.34	0.22	100.000	0.000
700	8	3.23	0.11	100.000	0.000
700	9	4.17	0.27	100.000	0.000
700	10	4.20	0.12	99.992	0.004
800	4	1.28	0.12	100.000	0.000
800	5	1.99	0.27	100.000	0.000
800	6	2.35	0.06	100.000	0.000
800	7	2.63	0.21	100.000	0.000
800	8	3.41	0.27	100.000	0.000
800	9	3.52	0.09	100.000	0.000
800	10	3.57	0.47	100.000	0.000
900	4	1.13	0.06	100.000	0.000
900	5	1.68	0.03	100.000	0.000
900	6	2.19	0.03	100.000	0.000
900	7	2.39	0.32	100.000	0.000
900	8	2.45	0.09	100.000	0.000
900	9	3.35	0.06	100.000	0.000
900	10	3.58	0.09	100.000	0.000
1000	4	0.85	0.03	100.000	0.000
1000	5	1.58	0.08	100.000	0.000
1000	6	1.75	0.01	100.000	0.000
1000	7	2.21	0.00	100.000	0.000
1000	8	2.83	0.13	100.000	0.000
1000	9	2.76	0.22	100.000	0.000
1000	10	3.21	0.13	100.000	0.000
1100	4	0.87	0.02	100.000	0.000
1100	5	1.29	0.05	100.000	0.000
1100	6	1.65	0.15	100.000	0.000
1100	7	2.23	0.18	100.000	0.000
1100	8	2.45	0.05	100.000	0.000

$I$	$K$	$R[\%]$	$\delta R[\%]$	$P[\%]$	$\delta P[\%]$
1100	9	2.84	0.28	99.988	0.006
1100	10	2.91	0.32	100.000	0.000
1200	4	0.74	0.00	100.000	0.000
1200	5	1.21	0.03	100.000	0.000
1200	6	1.47	0.12	100.000	0.000
1200	7	2.17	0.25	100.000	0.000
1200	8	2.40	0.11	100.000	0.000
1200	9	2.72	0.02	100.000	0.000
1200	10	2.83	0.25	100.000	0.000
1300	4	0.68	0.01	100.000	0.000
1300	5	1.15	0.10	100.000	0.000
1300	6	1.42	0.01	100.000	0.000
1300	7	1.74	0.05	100.000	0.000
1300	8	2.14	0.28	100.000	0.000
1300	9	2.26	0.18	100.000	0.000
1300	10	2.45	0.17	100.000	0.000

**Tabelle 10:** Precision und Recall, sowie deren Fehler, für den Cluster  $RLogL > 9.5$  in Abhängigkeit von  $I$  und  $K$ .

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
200	4	22.79	0.34	100.000	0.000
200	5	25.54	0.18	99.999	0.003
200	6	27.89	0.75	99.995	0.001
200	7	28.47	0.97	99.996	0.001
200	8	29.84	0.93	99.997	0.001
200	9	30.84	1.19	99.996	0.002
200	10	32.19	1.78	99.996	0.002
300	4	19.76	0.45	100.000	0.000
300	5	22.30	0.29	99.999	0.001
300	6	23.62	0.16	99.996	0.001
300	7	24.91	0.16	99.997	0.001
300	8	26.00	0.44	99.999	0.001
300	9	27.95	0.40	99.995	0.002
300	10	27.91	0.12	99.995	0.004
400	4	17.48	0.06	100.000	0.000
400	5	19.94	0.29	100.000	0.000
400	6	22.27	0.11	99.999	0.003
400	7	23.17	0.00	100.000	0.000
400	8	24.37	0.68	100.000	0.000
400	9	24.87	0.60	99.997	0.002
400	10	25.62	0.07	99.999	0.001
500	4	16.48	0.06	100.000	0.000
500	5	17.81	0.53	100.000	0.000
500	6	20.03	0.43	99.998	0.003
500	7	21.49	0.01	100.000	0.000
500	8	23.57	0.66	99.999	0.001
500	9	23.32	0.04	99.999	0.003
500	10	25.04	0.97	99.997	0.001
600	4	15.07	0.22	100.000	0.000
600	5	16.95	0.51	100.000	0.000
600	6	18.93	0.13	100.000	0.000
600	7	19.78	0.17	99.998	0.001
600	8	21.62	0.41	100.000	0.000
600	9	21.93	0.63	100.000	0.000

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
600	10	22.74	0.32	100.000	0.000
700	4	13.61	0.38	100.000	0.000
700	5	15.97	0.04	100.000	0.000
700	6	18.05	0.23	100.000	0.000
700	7	18.99	0.01	100.000	0.000
700	8	20.53	0.15	100.000	0.000
700	9	20.60	0.62	100.000	0.000
700	10	21.41	0.09	99.998	0.001
800	4	13.19	0.11	100.000	0.000
800	5	15.54	0.25	100.000	0.000
800	6	17.29	0.03	100.000	0.000
800	7	17.96	0.10	99.998	0.001
800	8	18.76	0.18	99.998	0.004
800	9	20.26	0.07	99.998	0.001
800	10	20.94	0.40	99.997	0.002
900	4	11.99	0.06	100.000	0.000
900	5	14.00	0.40	100.000	0.000
900	6	16.38	0.02	100.000	0.000
900	7	17.13	0.18	100.000	0.000
900	8	19.01	0.62	100.000	0.000
900	9	18.88	0.37	100.000	0.000
900	10	19.59	0.15	100.000	0.000
1000	4	11.89	0.00	100.000	0.000
1000	5	13.71	0.33	100.000	0.000
1000	6	15.34	0.19	100.000	0.000
1000	7	17.47	0.21	100.000	0.000
1000	8	18.05	0.21	100.000	0.000
1000	9	18.51	0.32	100.000	0.000
1000	10	19.75	0.11	100.000	0.000
1100	4	11.44	0.25	100.000	0.000
1100	5	13.36	0.11	100.000	0.000
1100	6	14.82	0.10	100.000	0.000
1100	7	16.47	0.25	100.000	0.000
1100	8	17.04	0.68	100.000	0.000

$I$	$K$	$R[\%]$	$\delta R[\%]$	$P[\%]$	$\delta P[\%]$
1100	9	18.03	0.21	100.000	0.000
1100	10	18.34	0.12	100.000	0.000
1200	4	10.75	0.16	100.000	0.000
1200	5	12.69	0.24	100.000	0.000
1200	6	14.14	0.17	100.000	0.000
1200	7	15.51	0.05	100.000	0.000
1200	8	16.66	0.30	100.000	0.000
1200	9	16.89	0.28	100.000	0.000
1200	10	18.34	0.21	100.000	0.000
1300	4	10.22	0.08	100.000	0.000
1300	5	12.49	0.08	100.000	0.000
1300	6	14.39	0.19	100.000	0.000
1300	7	15.14	0.17	100.000	0.000
1300	8	16.73	0.51	100.000	0.000
1300	9	17.02	0.12	100.000	0.000
1300	10	18.06	1.44	100.000	0.000

**Tabelle 11:** Precision und Recall, sowie deren Fehler, für den Cluster  $Zenit < 2.35rad$  in Abhängigkeit von  $I$  und  $K$ .

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
200	4	10.70	0.14	100.000	0.000
200	5	11.88	0.07	99.994	0.003
200	6	13.23	0.19	99.992	0.004
200	7	13.73	0.69	99.998	0.005
200	8	15.46	0.13	99.994	0.003
200	9	15.61	0.25	99.994	0.003
200	10	15.29	0.20	99.998	0.001
300	4	8.64	0.15	100.000	0.000
300	5	9.84	0.11	99.997	0.002
300	6	11.19	0.28	100.000	0.000
300	7	12.11	0.41	100.000	0.000
300	8	13.35	0.33	100.000	0.000
300	9	13.47	0.25	99.998	0.001
300	10	14.52	0.66	100.000	0.000
400	4	7.43	0.19	100.000	0.000
400	5	8.93	0.26	99.996	0.002
400	6	9.80	0.23	99.997	0.002
400	7	10.94	0.01	100.000	0.000
400	8	11.51	0.40	100.000	0.000
400	9	12.50	0.39	99.997	0.001
400	10	12.59	0.21	99.989	0.001
500	4	6.68	0.07	100.000	0.000
500	5	8.10	0.03	100.000	0.000
500	6	9.27	0.12	100.000	0.000
500	7	9.81	0.00	100.000	0.000
500	8	10.69	0.08	100.000	0.000
500	9	11.20	0.28	100.000	0.000
500	10	11.85	0.36	99.997	0.001
600	4	6.05	0.05	100.000	0.000
600	5	7.33	0.07	100.000	0.000
600	6	8.37	0.26	100.000	0.000
600	7	9.57	0.03	100.000	0.000
600	8	9.50	0.18	100.000	0.000
600	9	10.66	0.28	100.000	0.000

<i>I</i>	<i>K</i>	<i>R</i> [%]	$\delta R$ [%]	<i>P</i> [%]	$\delta P$ [%]
600	10	10.73	0.11	100.000	0.000
700	4	5.50	0.04	100.000	0.000
700	5	6.92	0.17	100.000	0.000
700	6	8.23	0.21	100.000	0.000
700	7	8.72	0.34	100.000	0.000
700	8	10.01	0.23	100.000	0.000
700	9	10.39	0.21	100.000	0.000
700	10	10.86	0.43	100.000	0.000
800	4	5.23	0.04	100.000	0.000
800	5	6.71	0.21	100.000	0.000
800	6	7.48	0.08	100.000	0.000
800	7	8.56	0.06	100.000	0.000
800	8	9.02	0.21	100.000	0.000
800	9	9.88	0.13	100.000	0.000
800	10	9.91	0.20	100.000	0.000
900	4	4.69	0.13	100.000	0.000
900	5	6.07	0.03	100.000	0.000
900	6	7.28	0.02	100.000	0.000
900	7	8.01	0.30	100.000	0.000
900	8	8.96	0.13	100.000	0.000
900	9	9.38	0.14	100.000	0.000
900	10	10.05	0.14	100.000	0.000
1000	4	4.53	0.11	100.000	0.000
1000	5	5.94	0.06	100.000	0.000
1000	6	6.97	0.07	100.000	0.000
1000	7	7.78	0.06	100.000	0.000
1000	8	8.23	0.15	100.000	0.000
1000	9	9.25	0.29	100.000	0.000
1000	10	9.42	0.13	100.000	0.000
1100	4	4.18	0.02	100.000	0.000
1100	5	5.56	0.16	100.000	0.000
1100	6	6.59	0.06	100.000	0.000
1100	7	7.52	0.03	100.000	0.000
1100	8	8.28	0.12	100.000	0.000

$I$	$K$	$R[\%]$	$\delta R[\%]$	$P[\%]$	$\delta P[\%]$
1100	9	8.94	0.01	100.000	0.000
1100	10	9.31	0.14	100.000	0.000
1200	4	3.96	0.13	100.000	0.000
1200	5	5.34	0.05	100.000	0.000
1200	6	6.46	0.06	100.000	0.000
1200	7	7.43	0.09	100.000	0.000
1200	8	8.00	0.33	100.000	0.000
1200	9	8.40	0.24	100.000	0.000
1200	10	9.48	0.17	100.000	0.000
1300	4	3.93	0.13	100.000	0.000
1300	5	5.13	0.03	100.000	0.000
1300	6	6.31	0.06	100.000	0.000
1300	7	7.14	0.10	100.000	0.000
1300	8	7.93	0.05	100.000	0.000
1300	9	8.36	0.05	100.000	0.000
1300	10	8.67	0.01	100.000	0.000

**Tabelle 12:** Precision und Recall, sowie deren Fehler, für den Cluster  $Zenit > 2.35rad$  in Abhängigkeit von  $I$  und  $K$ .

## Literatur

- [1] GALILEI, Galileo: *II Saggiatore (1623)*. Edition Nazionale, Bd. 6, Florenz 1896,
- [2] GANDHI, Raj ; QUIGG, Chris ; RENO, Mary H. ; SARCEVIC, Ina: *Ultrahigh-Energy Neutrino Interactions*. Astropart.Phys.5:81-110,1996
- [3] TIM RUHE, KATHARINA MORIK AND BENJAMIN SCHOWE: Data Mining on Ice, 2011
- [4] FLORIAN SCHERIAU: *Data-Mining fuer den IceCube Detektor*. 2009
- [5] DOLAG, Klaus ; GRASSO, Dario ; SPRINGEL, Volker ; TKACHEV, Igor: Constrained simulations of the magnetic field in the local Universe and the propagation of ultrahigh energy cosmic rays. In: *Journal of Cosmology and Astroparticle Physics* 2005 (2005), Nr. 01, 009. <http://stacks.iop.org/1475-7516/2005/i=01/a=009>
- [6] WOLFGANG WAGNER: *Design and Realisation of a new AMANDA Data Acquisition System with Transient Waveform Recorders*, Universität Dortmund, Diss., 2004
- [7] GREISEN, K.: End to the Cosmic-Ray Spectrum? In: *Physical Review Letters* 16 (1966), April, S. 748–750. <http://dx.doi.org/10.1103/PhysRevLett.16.748>. – DOI 10.1103/PhysRevLett.16.748
- [8] TYCE DEYOUNG: *Neutrino Astronomy with IceCube*. Mod.Phys.Lett.A24:1543-1557,2009
- [9] BRUCE A. REMINGTON, DAVID ARNETT, R. PAUL , DRAKE AND HIDEAKI TAKABE: Modeling Astrophysical Phenomena in the Laboratory with Intense Lasers. In: *Science* 28 (1999), May, S. 1488–1493
- [10] PROTHEROE, R. J.: *High Energy Neutrinos from Blazars*. 1996
- [11] RALPH ENGEL AND DAVID SECKEL AND TODOR STANEV: *Neutrinos from propagation of ultra-high energy protons*. Phys.Rev.D64:093010,2001
- [12] NAKAMURA, K. et a.: Review of Particle Physics. In: *Physical Review D* (2002). <http://pdg.lbl.gov>
- [13] GAISSER, T. K. ; HONDA, M.: Flux of Atmospheric Neutrinos. (2002). <http://dx.doi.org/10.1146/annurev.nucl.52.050102.090645>. – DOI 10.1146/annurev.nucl.52.050102.090645

- [14] WOSCHNAGG, Kurt: *Distances.i59.jpg*. <http://wiki.icecube.wisc.edu/index.php/File:Distances.i59.jpg>,
- [15] KOHAVI, Ron: *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. 1995
- [16] AL., Breiman et: *Classification and Regression Trees*. 1984
- [17] LEO BREIMAN AND E. SCHAPIRE: Random Forests. In: *Machine Learning*, 2001, S. 5–32
- [18] GAZIZOV, Askhat ; KOWALSKI, Marek P.: *ANIS: High Energy Neutrino Generator for Neutrino Telescopes*. *Comput.Phys.Commun.* 172 (2005) 203-213
- [19] D. HECK, J.N. Capdevielle G. Schatz T. T. J. Knapp K. J. Knapp: *CORSIKA: A Monte Carlo Code to Simulate Extensive Air Showers*. Forschungszentrum Karlsruhe Report FZKA 6019 (1998), 1998
- [20] BLOBEL, VOLKER AND LOHRMANN, ERICH: *Statistische und numerische Methoden der Datenanalyse*. 1. Teubner Verlag, 1998. – ISBN 3519032430
- [21] WIKI, IceCube: *rlogl*. [wiki.icecube.wisc.edu/index.php/Rlogl](http://wiki.icecube.wisc.edu/index.php/Rlogl), Juni 2010
- [22] SAMUEL ROBERT WHITEHEAD: *On the Properties of Ice at the IceCube Neutrino Telescope*, Department of Physics and Astronomy, University of Canterbury, Diplomarbeit, 2008
- [23] HONDA, M. AND KAJITA, T. AND KASAHARA, K. AND MIDORIKAWA, S. AND SANUKI, T.: Calculation of atmospheric neutrino flux using the interaction model calibrated with atmospheric muon data. In: *Phys. Rev. D* 75 (2007), Feb, 043006. <http://dx.doi.org/10.1103/PhysRevD.75.043006>. – DOI 10.1103/PhysRevD.75.043006
- [24] BUGAEV, E. V.; NAUMOV, V. A.; SINEGOVSKII, S. I.; ZASLAVSKAIA, E. S.: Prompt leptons in cosmic rays. In: *Nuovo Cimento C Serie 1 (ISSN 0390-5551)* 12 C (1989), Jan.-Feb., S. 41–73
- [25] KONDO, Yumi ; SALIBIAN-BARRERA, Matias ; ZAMAR, Ruben: *A robust and sparse K-means clustering algorithm*. 2012